PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

MAESTRÍA EN REDES DE COMUNICACIONES

INFORME FINAL DEL CASO DE ESTUDIO PARA UNIDAD DE TITULACIÓN ESPECIAL

TEMA:

"Análisis, consideraciones diseño y simulación a nivel de laboratorio de un sistema de Optimización de canal de ancho de banda para red WAN. Caso de estudio Unidad Educativa Eloy Alfaro de Santo Domingo de los Colorados"

Octubre, 30

Quito – 2017

AUTORÍA

Yo, Víctor René García Peña, portador de la cédula de ciudadanía No. 171395089-5, declaro bajo juramento que la presente investigación es de total responsabilidad del autor, y que se ha respetado las diferentes fuentes de información realizando las citas correspondientes. Esta investigación no contiene plagio y es resultado de un trabajo serio desarrollado en su totalidad por mi persona.

_____

Víctor René Garcia Peña
C.C. 171395089-5

Contenido

1. INTRODUCCIÓN 1
2. JUSTIFICACIÓN 3
3. ANTECEDENTES 4
4. OBJETIVOS 5
4.1. Objetivo General 5
4.2. Objetivos Específicos 5
5. ESTADO DEL ARTE DE LOS SISTEMAS DE OPTIMIZACIÓN DE ANCHO DE BANDA WAN 6
5.1. La problemática del ancho de banda WAN 6
5.2. Protocolos más usados en la Optimización WAN 8
5.2.1. TCP 8
5.2.2. HTTP 9
5.2.3. HTTP/S 9
5.2.4. SSL 10
5.2.5. CIFS 11
5.2.6. FTP 11
5.3. Técnicas de optimización WAN 12
6. RETOS PARA LA IMPLEMENTACIÓN DE SOLUCIONES DE OPTIMIZACIÓN DE ANCHO DE BANDA WAN 15
6.1. Descripción de la aceleración específica de aplicaciones 16
6.2. Almacenamiento en caché específico para aplicaciones 17
6.3. Ventajas del almacenamiento en caché específico para aplicaciones 18
6.4. Validación del caché y frescura del contenido 19
6.4.1. CIFS 23
6.4.2. HTTP 28
6.4.3. Media Streaming: RTSP, HTTP, y Flash 29
6.4.4. Aplicaciones de bases de datos basadas en Web 31
7. LIMITACIONES DE LOS ENLACES DE DATOS WAN. 36
7.1. Limitaciones del Transporte de Protocolo 37
7.2. Fundamentos del Protocolo de control de transmisión 39
7.2.1. Servicios orientados a conexión 40
7.2.2. Garantizando la Entrega 42

7.2.3. Descubriendo del Ancho de banda 46
7.2.4. Inicio lento de TCP 47
7.2.5. Mecanismos para evitar la congestión TCP 48
8. ANÁLISIS DE LAS HERRAMIENTAS DE OPTIMIZACIÓN DISPONIBLES EN EL MERCADO 50
8.1. Entorno del pruebas de las herramientas de optimización 50
8.2. Pruebas de Rendimiento 53
8.3. Tráfico HTTP 54
8.4. Tráfico HTTPS 54
8.5. Tráfico de correo electrónico 55
8.6. Tráfico de Voz sobre IP VoIP 56
8.7. Gestión de Tráfico 57
8.8. Visibilidad 58
8.9. Facilidad de uso 58
9. Elección de la herramienta de optimización para el ambiente de Laboratorio. 60
10. ARQUITECTURA DE LA SOLUCIÓN DE OPTIMIZACIÓN DE ANCHO DE BANDA WAN PARA EL AMBIENTE DE LABORATORIO 63
10.1. Parámetros de lógicos del laboratorio 65
11. DESARROLLO DE LA SOLUCIÓN EN LABORATORIO 66
11.1. Configuración del ambiente de laboratorio 66
11.2. Arquitectura lógica del ambiente de laboratorio 67
11.3. ESCENARIO DE LABORATORIO 68
11.3.1. CONSIDERACIONES DE LABORATORIO 68
11.3.2. Transferencia de archivos en frío mediante CIFS/SMBv1 71
11.3.3. Descarga de información desde un servidor HTTP 73
11.3.4. Acceso múltiple de usuarios a una Página Web de la Intranet 74
11.4. Análisis de Resultados 76
12. ANÁLISIS DE COSTOS DE LA SOLUCIÓN DE INFRAESTRUCTURA OPTIMIZACIÓN DE ANCHO DE BANDA WAN DE RIVERBED STEELHEAD 78
13. CONCLUSIONES Y RECOMENDACIONES 80
14. BIBLIOGRAFÍA 82

Listado de Figuras

Figura 1: Uso de Internet en el Ecuador por Zonas 3
Figura 2: Modelo TCP/IP 8
Figura 3: Validación de Cache con Coherencia No estricta 22
Figura 4: Validación de Cache con Coherencia estricta 23
Figura 5: Servidor de caché para objetos de Lógica Parcial 27
Figura 6: División del streaming para minimizar el consumo de Ancho de Banda WAN 30
Figura 7: Despliegue de aceleradores para aplicaciones Web y de Base de Datos 34
Figura 8: Aceleración específica para aplicaciones de Base de Datos 35
Figura 9: Aceleración de aplicaciones y jerarquía de optimización de WAN 39
Figura 10: Socket TCP 40
Figura 11: Establecimiento de la conexión TCP 42
Figura 12: Buffer TCP entre la red y las aplicaciones 43
Figura 13: Operación TCP 44
Figura 14: Gestión de retransmisión TCP 46
Figura 15: ScoreCard de los productos de optimización 52
Figura 16: Comparación de tráfico HTTP [8] 54
Figura 17: Comparación de tráfico de Correo Electrónico [8] 55
Figura 18: Comparación de tráfico de Voz sobre IP [8] 56
Figura 19: Cuadrante de Gartner para Optimización WAN 61
Figura 20: Ambiente físico para Simulación 63
Figura 21: Modelos de Equipamiento SteelHead [9] 64
Figura 22: Opciones de Configuración de la Herramienta WANem 65
Figura 23: Comando para configuración ruteo 66
Figura 24: Configuración de Equipo SteelHead Virtual [10] 67
Figura 25: Estado de Salud del Equipo Virtual 67
Figura 26: Diagrama de Red del Laboratorio 68
Figura 27: Prueba de Conectividad 69
Figura 28: Verificación de la velocidad del Canal dentro de la consola SteelHead 69
Figura 29: Verificación de conectividad entre equipos 70
Figura 30: Informe de “Current Connections” 70
Figura 31: Reinicio de servicios de Optimización 71
Figura 32: Información de Optimización de conexiones 72
Figura 33: Información de tráfico optimizado 72
Figura 34: Optimización de Canal 73
Figura 35: Configuración de Apache JMeter 74
Figura 36: Captura de pantalla de la optimización WAN de la herramienta Riverbed 75
Figura 37: Optimización de ancho de Banda en prueba de usuarios múltiples 76

Listado de Tablas

Tabla 1: Técnicas de Optimización WAN 14
Tabla 2: Información de intercambio durante la conexión TCP 41
Tabla 3: Soluciones de Optmización WAN – Disponibilidad de Licencia de Demostración..... 61
Tabla 4: Soluciones de Optmización WAN – Disponibilidad de Licencia de Demostración..... 65
Tabla 5: Pruebas de envío de información 71
Tabla 6: Costo del Enlace proyectado a 3 y 5 años..... 78
Tabla 7: Costo del Enlace proyectado a 3 años..... 79
Tabla 8: Costo del Enlace proyectado a 5 años..... 79

1. INTRODUCCIÓN

El crecimiento y despliegue de los sistemas informáticos en la actualidad han contribuido a que las redes de datos se consideren una de las principales herramientas en las empresas para acceder a servicios y ejecutar sus operaciones diarias. Muchas operaciones del negocio de las empresas hoy en día hacen uso de la red de la empresa, existiendo incluso escenarios en los que se extiende la cobertura de la red a oficinas remotas o sucursales mediante la incorporación de enlaces WAN, sin embargo muchas de las aplicaciones y servicios han sido desarrollados para su uso dentro de una LAN, por lo que estos nuevos esquemas de expansión conllevan un impacto en el desempeño de las aplicaciones. Por esto motivo el estudio y análisis de redes extendidas han experimentado un incremento en la atención por parte de clientes y usuarios, al requerir de mayores capacidades y desempeño que puedan satisfacer las necesidades de los usuarios.

En el caso de la Unidad Educativa “Eloy Alfaro” en los recientes años se han implementado servicios de correo electrónico, acceso de información de estudiantes vía web, compartición de recursos educativos a través de la red, entre algunos de los importantes siendo de gran importancia la comunicación con la Dirección Distrital de Educación 23D02 y los Sistemas del Ministerio de Educación, a través de un enlace WAN con una capacidad de 1.544 Mbps y por el cuál se conectan usuarios que continuamente han presentado quejas de lentitud e incapacidad de realizar sus tareas diarias. Ante esta problemática se ha planteado la necesidad de incrementar las capacidades del enlace de comunicaciones, siendo una alternativa la de usar optimizadores de Ancho de Banda que permitan optimizar los recursos actuales y representen un ahorro en la Unidad Educativa.

El presente Caso de Estudio busca estudiar el desempeño de una red tipo WAN para mejorar su desempeño mediante el uso de optimizadores, enfocándose en la necesidad de la Unidad Educativa “Eloy Alfaro” y determinar según un análisis de mercado la mejor solución que pueda cumplir con sus necesidades y representar un ahorro económico a largo plazo.

Dentro de las organizaciones el tráfico de tipo recreacional o no esencial muchas veces puede ocupar cerca del 50% del total del canal WAN [2]. Esto enfrenta a los administradores de TI a realizar cambios en su infraestructura para enfrentar la necesidad de satisfacer la demanda de toda la organización. Estos cambios en muchas ocasiones se enfocan en la restricción al acceso de aplicaciones, servicios y contenidos, sin embargo en la actualidad estas soluciones suelen ser muy poco eficientes ya que los usuarios requieren de los contenidos para la ejecución de sus tareas. Esta problemática ha dado paso a un mercado de herramientas de optimización WAN que se enfocan la

utilización de equipos intermedios entre la LAN y la WAN que buscan reducir la cantidad de información que viaja a través del enlace de datos WAN para lo cual usan técnicas de deduplicación de la información, optimización TCP, Caching de información, compresión de información previo a su envío, entre algunas que se verán más adelante como parte de este trabajo.

En los primeros capítulos de este trabajo se analizará los antecedentes de la Unidad Educativa, su problemática y el escenario en el que se busca implementar la solución que se armará en laboratorio. Adicionalmente se revisará los conceptos de estas herramientas y cómo interactúan en los ambientes de las organizaciones.

Más adelante se analizará los desafíos que se deben superar para la implementación de soluciones de optimización de ancho de banda y los requisitos previos que debe sortear la organización para implementar exitosamente este tipo de soluciones. Posterior a este se analizará los conceptos de optimización y las técnicas utilizadas para entregar una solución a las instituciones que buscan implementar este tipo de soluciones.

En los capítulos subsiguientes se explicará los escenarios en los que se puede implementar estas soluciones haciendo un análisis del escenario actual de la Unidad Educativa "Eloy Alfaro". Como parte de este trabajo se incluirá un análisis comparativo de las soluciones existentes en el mercado y justificará la elección de la solución para el ambiente de laboratorio.

Ya dentro del escenario de laboratorio se presentará la arquitectura de la solución de optimización de ancho de banda WAN a manera general para luego hacer una aproximación al ambiente de la organización y aterrizar en un diseño para este ambiente que se reproducirá en el laboratorio. Dentro del análisis de laboratorio se crearán los escenarios, bancos de pruebas y se analizará los resultados para determinar la factibilidad del ambiente en un escenario productivo.

Finalmente se realizará un análisis general de la solución y se realizará un análisis final de factibilidad para la Unidad Educativa "Eloy Alfaro", realizando las mejores recomendaciones para que el proyecto represente una opción atractiva para la organización. Se cerrará esta sección con las conclusiones y recomendaciones finales del trabajo realizado.

El desarrollo de este trabajo analiza la problemática de la optimización de recursos tecnológicos limitados en la Unidad Educativa "Eloy Alfaro" que posee enlaces WAN hacia otras dependencias del Distrito de Educación para el acceso de servicios de correo electrónico, FTP, INTRANET, etc. Para esto cubrirá un análisis, consideraciones diseño y recomendaciones en laboratorio y de esta manera

si las prestaciones de la solución permiten optimizar los recursos y es un escenario viable desde el punto de vista técnico para la institución.

2. JUSTIFICACIÓN

El acceso a Internet y redes de contenidos ha experimentado un crecimiento constante a nivel mundial y específicamente en el Ecuador según cifras del INEC [3] ha tenido una tendencia al alza desde el año 2012 hasta la actualidad situándose en el 2016 en cifras de acceso al Internet en el área urbana del 44.6% y en áreas rurales del 16,4%. Esto se traduce en la generación de más tráfico sobre las redes de datos que deben soportar la demanda que generan los usuarios, en el mismo estudio del INEC y según la Figura 1 se observa que el acceso a Internet a nivel nacional es de cerca del 55.6%. Si bien estos valores no se pueden correlacionar con el acceso a las redes de datos empresariales nos da una idea de cómo abarca la tecnología la vida de las personas y a su vez implica retos hacia las empresas para satisfacer la demanda de sus usuarios.

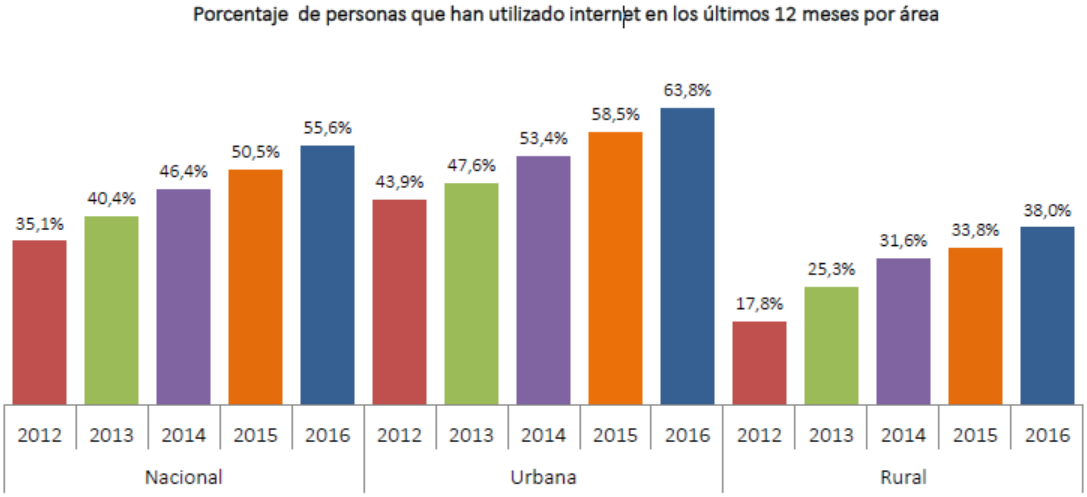


Figura 1: *Uso de Internet en el Ecuador por Zonas*
Fuente: *http://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/TIC/2016/170125.Presentacion_Tics_2016.pdf*

Bajo este esquema se ha planteado el requerimiento de ampliar la capacidad de las redes actuales, sin embargo esta necesidad va de la mano con el incremento de inversión en tecnologías, enlaces de datos y capacidades de los actuales servicios de tecnología, por ese motivo se ha visto la necesidad plantear soluciones que optimicen los recursos de las empresas para que puedan afrontar el requerimiento de mejores capacidades sin una inversión grande de recursos económicos.

Por este motivo se ha centrado este caso de estudio en la optimización de recursos de enlaces de datos WAN con el objetivo de ofrecer una solución que reduzca los costos al corto y largo plazo y represente una opción viable para la Unidad Educativa "Eloy Alfaro". En muchas ocasiones la solución ante limitaciones de anchos de banda WAN se plantea en el incremento de la capacidad de dichos enlaces, sin embargo si se los números del crecimiento de usuarios en el mismo estudio del INEC se puede observar que a nivel nacional en el Ecuador en el 2016 cerca del 55.6% de persona accedieron a Internet y las cifras han venido creciendo desde el 2012 lo que marca una tendencia del comportamiento de las personas hacia los servicios basados en redes de datos.

Finalmente para el desarrollo de este caso de estudio se ha visto la necesidad de implementar una solución que se probará a nivel de laboratorio que optimice los enlaces de datos y no requiera de actualización de la infraestructura existente y que adicionalmente no represente una carga adicional de administración y conocimiento técnico específico sobre los componentes que formen parte de la solución planteada.

3. ANTECEDENTES

La Unidad Educativa "Eloy Alfaro" es una institución pública que presta sus servicios a la comunidad de Santo Domingo en el sector de Las Palmas, como muchas instituciones educativas han debido desarrollar procesos de modernización y generación de ofertas de servicios de tecnología para sus usuarios y estudiantes. Durante el año 2014 se implementaron servicios de correo electrónico, página web institucional, repositorio de material educativo (documentos e imágenes) para uso de profesores y personal administrativo, adicionalmente se implementó un servicio web para la publicación de calificaciones de los estudiantes para que los padres de familia puedan seguir de cerca el desempeño de sus hijos. En este contexto la institución maneja oficinas administrativas y de tecnología fuera del claustro escolar, principalmente por temas de seguridad del equipamiento y facilidades técnicas de alojamiento.

Por esta topología ha sido necesario en todo momento contar con el servicio de enlace de datos entre en el Data Center de la institución y la oficinas administrativas. Actualmente se cuenta con un enlace de 1.544 Mbps provisto por la empresa CNT y que según información de la misma empresa tiene una saturación cerca al 90%, motivo por el cual la Unidad Educativa "Eloy Alfaro", ha venido analizando la posibilidad de realizar un incremento de capacidad del enlace o de limitar el uso de determinados servicios. Ante esta problemática se ha planteado a la institución la posibilidad de

realizar una inversión a nivel de equipamiento que optimice su recurso de enlace de datos y optimice su capacidad actual.

De una inspección inicial se ha determinado que la mayor necesidad del cliente es la de mantener documentación que generan los usuarios, como reportes y documentación digitalizada de alumnos y profesores. Ante esto se ha visto que el mercado existen varias soluciones que optimizan los canales WAN y puede representar una opción viable y que a largo plazo puede representar un ahorro para la institución. Más adelante en este trabajo se analizará las opciones que existen en el mercado y se realizará un análisis de la mejor opción para la Unidad Educativa "Eloy Alfaro".

4. OBJETIVOS

4.1. OBJETIVO GENERAL

Realizar un análisis, establecer consideraciones de diseño, y presentar una simulación a nivel de laboratorio de un sistema de optimización de ancho de banda WAN, para la Unidad Educativa "Eloy Alfaro".

4.2. OBJETIVOS ESPECÍFICOS

a) Presentar el estado del Arte de los sistemas de optimización de ancho de banda WAN.
b) Entender las barreras para la implementación de soluciones de Optimización de Ancho de Banda.
c) Entender las limitaciones de los enlaces de Datos WAN.
d) Hacer un análisis comparativo de las principales soluciones de optimización de ancho de banda WAN del mercado.
e) Explicar la arquitectura de la solución de optimización de ancho de banda WAN.
f) Presentar y analizar un diseño de solución de optimización de ancho de banda WAN dado los requerimientos planteados en el escenario particular del presente caso de estudio.
g) Realizar una simulación del diseño propuesto a nivel de laboratorio virtual, analizar los resultados y en base a ellos emitir recomendaciones.
h) Realizar un análisis de costos de la solución de infraestructura optimización de ancho de banda WAN usada para la simulación en laboratorio.

5. ESTADO DEL ARTE DE LOS SISTEMAS DE OPTIMIZACIÓN DE ANCHO DE BANDA WAN

Las técnicas de optimización WAN son un conjunto de técnicas y estrategias que mejoran la eficacia al momento de enviar información a través de enlaces de comunicación dedicados (WAN). Las organizaciones han desarrollado redes de datos por las que envían información de los usuarios que en muchas ocasiones deben viajar grandes distancias físicas para llegar a su ubicación final. Por lo general las empresas y organizaciones integran un número importante de aplicaciones que trabajan sobre la red empresarial, de manera que el ancho de banda se debe compartir entre todas los servicios que posee la organización [1][2]. Para esto es necesario entender que los escenarios más comunes de topologías WAN son:

- Oficina Central – Sucursales
- Centro de Datos A – Centro de Datos B.

De estas dos opciones la de Oficina Central – Sucursales es la más común y la que supone una menor distancia entre ambas ubicaciones y por lo general consumen menos ancho de banda, soportan más conexiones simultáneas y de menor tamaño, así como conexiones de corta duración con una gran cantidad de protocolos que viajan a través del enlace de datos. Este tipo de enlaces son usados para aplicaciones de las organizaciones como correo electrónico, sistemas de gestión de contenido, acceso a bases de datos y consulta de servicios web.

Por otro lado existen las conexiones del tipo Data Center A – Data Center B, que conllevan el uso de enlaces de comunicaciones de mayor capacidad, recorren distancias mayores y requieren de menores conexiones pero generan una mayor cantidad de información con tiempos de conexión mayores (mayores volúmenes de transferencia de información). En ambos escenarios es importante balancear correctamente la capacidad de los enlaces entre las diferentes necesidades para no provocar una afectación al negocio.[6]

5.1. LA PROBLEMÁTICA DEL ANCHO DE BANDA WAN

Cuando las aplicaciones cruzan a través de los enlaces WAN y tienen un desempeño pobre caracterizado por la lentitud, los administradores de TI generalmente asumen que es el resultado de la capacidad del ancho de banda. Sin embargo este concepto es poco acertado y generalmente el desempeño de la WAN es producto de muchos factores como los siguientes:

a) Tiempo de Respuesta de las aplicaciones

En ambientes distribuidos de usuarios y aplicaciones, los usuarios y las oficinas remotas generalmente están ubicados en locaciones remotas, siendo el esquema de acceso a los recursos de TI de manera centralizada en un Centro de Datos que concentra todos los servicios. La distancia entre el Centro de Datos y las oficinas remotas conlleva a un degrado en el desempeño de las aplicaciones. Los paquetes de datos toman un determinado tiempo en viajar desde su origen al destino lo cual se considera la latencia propia del canal, que combinada con un ancho de banda escaso y aplicaciones web ineficientes resultan en transferencias de información interminables. [1]

b) Protocolos ineficientes

Cuando los protocolos de comunicaciones envían información en pequeños paquetes, de manera secuencial a través de la red se conoce como Protocolos Habladores o "Chatty Protocols". Para evitar la pérdida de paquetes, estos protocolos por lo general dividen la información en bastantes paquetes pequeños antes de ser enviados a la red. [6]

c) Retardo

La latencia es definida como el intervalo de tiempo entre el origen y el destino. Al incrementar el ancho de banda no se soluciona los problemas de latencia y de Protocolos ineficientes. Para solucionar esta problemática se puede optar por incluir mejoras a nivel de TCP como el manejo del tamaño de las tramas o re-ubicación de las aplicaciones para que estén más cerca de su punto final de acceso.

La optimización WAN ha sido tema de extensa investigación académica casi desde el inicio de las redes WAN. A inicios de los 2000s, la investigación a nivel de los sectores privados y público se enfocaron en mejorar el throughput de TCP de extremo a extremo, y el objetivo de las primeras soluciones propietarias de optimización WAN se enfocaron en esta problemática. Sin embargo, en años recientes, el rápido crecimiento de los datos digitales, y las necesidades crecientes de almacenarlo y protegerlo, han creado un mercado de soluciones de optimización que van desde soluciones de código abierto hasta soluciones propietarias. [4]

Algunas estrategias que se incluyen en la optimización de enlaces WAN incluye técnicas de deduplicación, compresión, manejo de caché, optimización de la latencia, etc. y se analizarán más adelante. Todo este conjunto de técnicas se aplican a un conjunto de protocolos que desde el nacimiento de las redes de datos presentan sus propios retos y se analizan a continuación.

5.2. PROTOCOLOS MÁS USADOS EN LA OPTIMIZACIÓN WAN

5.2.1. TCP

TCP (Protocolo de Control de Transmisión) es un protocolo orientado a la conexión full-dúplex que provee un circuito virtual totalmente confiable para la transmisión de información entre dos aplicaciones. TCP garantiza que la información enviada llegue hasta su destino sin errores y en el mismo orden en que fue enviada.

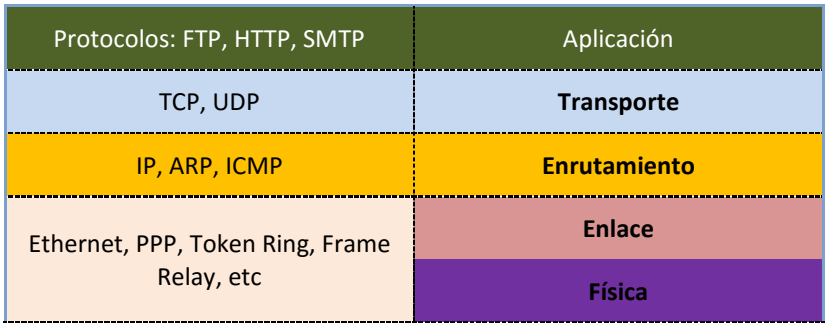


Figura 2: Modelo TCP/IP
Fuente: http://www.newdevices.com/tutoriales/modelo-tcpip/2.html

En la Figura 2 se describe el proceso del Modelo TCP/IP. TCP es un protocolo orientado a conexiones, los datos son transmitidos en segmentos. "Orientado a Conexiones" significa que se debe establecerse una conexión antes de que el host intercambie datos. La eficacia del protocolo se basa en la asignación de un número de secuencia a cada segmento transmitido. Se utiliza una confirmación para afirmar que los datos fueron recibidos por el otro servidor. Para cada segmento enviado, el servidor que recibe debe regresar una confirmación (acknowledgment o ACK) en un periodo específico de bytes recibidos. Si el ACK no es recibido, los datos son retransmitidos. [4][5]

TCP utiliza comunicaciones de flujo de bytes, donde los datos dentro del segmento TCP son tratados como una secuencia de bytes sin límites de registro o de campo. Un puerto TCP proporciona una localización específica para entregar los segmentos TCP. Los números de puertos por debajo de 1024 son puertos que se designan como "conocidos" y están asignados por la IANA (Autoridad de Número

Asignados de Internet), que en la actualidad está representada como ICANN (La Corporación de Internet para la Asignación de Nombres y Números).

TCP/IP (Transmisión Control Protocol / Internet Protocol) “es un grupo de protocolos estándares de la industria diseñados para redes. Se ha convertido en el protocolo más popular debido a que es utilizado por Internet y está muy extendido en los sistemas operativos. TCP/IP se ha convertido en el conjunto de protocolos de red disponible más adaptable por el medio del cual se puede trabajar casi en cualquier medio de Red, Hardware y Sistema Operativo existente, desde una pequeña LAN de grupo de trabajo, hasta la conexión de millones de sistemas que componen la propia Internet.”[7]

5.2.2. HTTP

HTTP (Protocolo de Transferencia de Hipertexto), es un protocolo que trabaja a nivel de aplicación y es usado para transferir información entre sistemas de forma clara y rápida. HTTP ha sido utilizado por la World Wide Web desde 1990. HTTP es el protocolo encargado de dar vida a Internet permitiendo que los clientes y servidores se pueden comunicar. Actualmente se recomienda que se utilice el protocolo HTTP 1.1, debido a que la versión 1.0 está desapareciendo lentamente. La nueva versión de HTTP está recogida dentro de la RFC 2068 de enero de 1997.[5][7]

5.2.3. HTTP/S

HTTPS (Protocolo Seguro de Transferencia de Hipertexto) es la combinación de los protocolos HTTP y protocolos criptográficos. HTTPS es utilizado para conseguir conexiones seguras en la WWW, habitualmente para transacciones de pagos o cada vez que se intercambie información importante en internet, como por ejemplo claves, etc. Es así que la información importante, en el caso de ser interceptada por un extraño, estará cifrada y no podrá ser legible.

El nivel de protección que ofrece HTTPS depende de muchos factores como la corrección de la implementación del navegador web, del software y de los algoritmos criptográficos soportados. El protocolo fue creado por Netscape Communications en 1994 para su navegador Netscape Navigator, para poder diferenciar una comunicación o página web segura.

Originalmente HTTPS sólo utilizaba la encriptación SSL (Capa de Conexión Segura), luego reemplazado por TLS (Seguridad en la Capa de Transporte). HTTPS fue adoptado como estándar web por el grupo IETF (Grupo Especial sobre Ingeniería de Internet) tras la publicación del RFC 2818 en mayo de 2000. [5][7]

5.2.4. SSL

Este protocolo es el que se encarga de proporcionar autenticación y privacidad de la información entre los extremos de Internet utilizando herramientas criptográficas. Normalmente, el servidor es autenticado mientras que el cliente se mantiene sin autenticar; para la autenticación mutua se requiere de un despliegue de infraestructura de claves públicas (PKI) para los clientes.

Fases básicas de SSL:

- Negociar entre las partes el algoritmo que se usará en la comunicación.
- Intercambio de claves públicas y autenticación basada en certificados digitales.
- Encriptación del tráfico basado en cifrado simétrico.

SSL se ejecuta en una capa entre los protocolos de aplicación como HTTP, SMTP (Protocolo Simple de Transferencia de Correo), NNTP (Protocolo para la Transferencia de Noticias en Red) y sobre el protocolo de transporte TCP, que forma parte de la familia de protocolos TCP/IP. Aunque pueda proporcionar seguridad a cualquier protocolo que use conexiones de confianza (tal como TCP), se usa en la mayoría de los casos junto a HTTP para formar HTTPS. [5][7]

El protocolo SSL/TLS cuenta con muchas aplicaciones en la actualidad. La mayoría son versiones seguras de programas que emplean diferentes protocolos. Hay versiones seguras de servidores y clientes de protocolos como http, nntp, imap, pop3, etc. También existen diferentes productos clientes y servidores que pueden proporcionar cifrado SSL de forma nativa, pero de igual manera existen muchos que aún no lo permiten.

El tráfico SSL representa un porcentaje cada vez mayor del tráfico total sobre enlaces de red WAN. Así pues, el soporte de SSL en dispositivos de optimización de redes WAN será cada vez más significativo para las empresas que quieren conservar el tráfico consolidado además de disminuir el tamaño de sus enlaces WAN. Algunas compañías como, Blue Coat, Certeon y Riverbed Technology, brindan aceleración de SSL en sus dispositivos. Estos dispositivos son colocados en ambos extremos de los enlaces de red y efectúan varias funciones que sirven para acelerar las transacciones. Esto incluye optimizar las sesiones TCP (Protocolo del Control de Transmisión), hacer cumplir las políticas de Calidad de Servicio - QoS o la optimización de protocolos. Sin el soporte de SSL, cuando el tráfico SSL llega a estas cajas, se limitan a utilizar la optimización de TCP y QoS. [5][7]

5.2.5. CIFS

CIFS (Protocolo de Sistema de Archivo Común de Internet), es una versión perfeccionada de Bloque de Mensaje de Servidor de Microsoft (SMB), es el camino estándar en el cual los usuarios de ordenador comparten archivos a través de la intranet e Internet. El CIFS permite la colaboración sobre Internet, definiendo un protocolo de acceso de archivo remoto que es compatible con el modo de uso que comparten datos sobre discos locales y conectan una red de servidores de ficheros. El CIFS funciona sobre TCP/IP y utiliza el Servicio de Nombramiento de Dominio (DNS) global de Internet para la adaptabilidad, y es optimizado para soportar una menor velocidad que las conexiones de acceso telefónico (dial-up) sobre Internet. CIFS puede ser enviado sobre una red a dispositivos remotos que usan los paquetes de nuevos directores. El nuevo director también usa el CIFS para hacer peticiones a la pila del protocolo del ordenador local. [6][7]

Microsoft denominó a (CIFS) como SMB en 1998 y añadió más características, que incluyen soporte para enlaces simbólicos, enlaces duros (hard links), y mayores tamaños de archivo. Hay características en la implementación de SMB de Microsoft que no son parte del protocolo SMB original. También existe Samba, que es una implementación libre del protocolo SMB con las extensiones de Microsoft. Funciona sobre sistemas operativos GNU/Linux y en otros UNIX. La característica más importante que ofrece CIFS es el acceso de archivos con integridad. CIFS apoya el juego habitual de operaciones de archivo; abrir, cerrar, leer, escribir y buscar. CIFS también soporta archivos, registro de bloqueo y desbloqueo. CIFS permite a múltiples clientes tener acceso y actualizar el mismo archivo al mismo tiempo previniendo conflictos, proporcionando el intercambio de archivos y el bloqueo de archivos. [6][7]

5.2.6. FTP

FTP (Protocolo de Transferencia de Archivos), define la manera en que los datos deben ser transferidos a través de una red TCP/IP. FTP es un protocolo estándar y se lo describe en el RFC 959. La copia de ficheros de una máquina a otra es una de las operaciones más frecuentes. La transferencia de datos entre cliente y servidor puede producirse en cualquier dirección. El cliente puede enviar o pedir un fichero al servidor. [6][7]

FTP emplea dos conexiones: la primera para el *login* y la segunda para gestionar la transferencia de datos. Debido a que se necesita hacer un *login* en el host remoto, el usuario deberá tener un nombre de usuario y un *password* para acceder a ficheros y a directorios. El usuario que inicie la conexión

asume la función de cliente, y el host remoto adopta la función de servidor. El objetivo del protocolo FTP es:

- Permitir que equipos remotos puedan compartir archivos.
- Permitir la independencia entre los sistemas de archivo del equipo del cliente y del equipo del servidor.
- Permitir una transferencia de datos eficaz.

FTP está contenido dentro del modelo cliente-servidor, es decir, un equipo envía órdenes (cliente) y el otro espera solicitudes para llevar a cabo acciones (servidor).

Durante la conexión de FTP, se encuentran abiertos dos canales de transmisión:

- Un canal de comandos (canal de control).
- Un canal de datos.

5.3. TÉCNICAS DE OPTIMIZACIÓN WAN

Las técnicas de optimización WAN están constituidas por un marco de tecnologías que mejoran la experiencia de las aplicaciones en la red, y hacen un mejor uso de los limitados recursos de la red. En algunos casos, la experiencia del usuario simplemente debe mantenerse mientras ocurren otros cambios. Son varios los mecanismos disponibles para la optimización de la WAN, que van desde las tecnologías que permiten escalabilidad horizontal (la capacidad de crear clústers con varios dispositivos, en lugar de escalabilidad vertical que requiere más potencia en cada uno de los dispositivos), la carga del servidor compartido, avanzadas tecnologías de compresión, encaminamiento dinámico para colocar tráfico en el mejor camino, reducción de replicación, flujo inteligente, entre otros. En la Tabla 1, existen varias técnicas que se pueden utilizar y que se describen según el alcance y características que presentan. [6][7]

Técnica	Definición

Deduplicación	Esta función disminuye las transmisiones de datos, eliminando la transferencia de datos redundantes a través de la WAN enviando referencias en vez del dato completo.
Compresión	La Compresión es una técnica para reducir tamaño de datos con el fin de ahorrar espacio o el tiempo de transmisión. Compresión de datos o fuente de codificación es el proceso donde la información se codifica mediante una codificación única, las estrategias es que utilizan menos bits u otras unidades portadoras de información. Se basa en patrones de datos que puede ser representado más eficientemente.
Optimización de latencia	Puede incluir refinamientos de TCP como escalado del tamaño de ventana, confirmación selectiva, algoritmos de control de congestión de Capa 3 e incluso estrategias de ubicación en las que la aplicación está colocada cerca del punto final para reducir latencia. En algunas implementaciones, el optimizador local WAN contestará localmente las peticiones del cliente en vez de enviar la petición al servidor remoto para apalancar los mecanismos de *write-behind* y *read-ahead* para reducir la latencia WAN.
Caching/proxy	Poner datos en caches locales; se basa en la predicción del comportamiento humano, que accede múltiples veces a la misma información
Corrección de error adelantada	Esta técnica se utiliza para el control de errores en una transmisión de datos a través de canales de comunicación poco fiables o ruidoso. La idea central es que el remitente codifica su mensaje en una forma redundante mediante el uso de un código de corrección de errores (ECC). De esta forma se mitiga la pérdida de paquetes añadiendo otro paquete de recuperación de pérdida - por cada "N" paquetes que está enviado, y esto reduce la necesidad de retransmisiones en enlaces WAN congestionados y que presentan errores
Engaño de protocolo	Empaqueta múltiples requerimientos de aplicaciones similares, en uno solo requerimiento. Puede incluir también limpiar protocolos como CIFS.
Conformado de tráfico	Controla el flujo de datos de aplicaciones concretas. Esto da flexibilidad a operadores de red y administradores para decidir qué aplicaciones toman precedencia sobre la WAN. Un caso de uso común de conformado de tráfico es que para impedir que un protocolo o aplicación acaparen el ancho de banda disponible en un enlace sobre otros protocolos considerados más importantes para el administrador/negocio. Algunos dispositivos de aceleración WAN son capaces de dar forma al tráfico con granularidad mucho más allá de los dispositivos de red tradicional.
Ecualización	Hace suposiciones en lo que necesita prioridad inmediata basada en el uso de datos. Ejemplos de uso de igualación puede incluir

	conexiones de Internet abiertas y no reguladas y túneles VPN congestionados.
Limitación de conexiones	Impide bloqueos de los puntos de acceso debido a ataques de denegación de servicio o conexiones *peer to peer*. Funciona mejor para enlaces de acceso del Internet abierto, aunque puede ser utilizado en Optimizadores WAN.
Limitación de tasa de transferencia	Impide que un usuario consiga más de una cantidad fija de datos.

Tabla 1: Técnicas de Optimización WAN [5][7]

6. RETOS PARA LA IMPLEMENTACIÓN DE SOLUCIONES DE OPTIMIZACIÓN DE ANCHO DE BANDA WAN

Los optimizadores WAN generalmente permiten la aceleración del envío de tráfico WAN mediante dos mecanismos, el primer mecanismo es independiente de la aplicación y se enfoca en optimizar directamente la WAN, este mecanismo se enfoca en superar y corregir las condiciones en las que se desarrolla la WAN es decir busca entregar mecanismos que mejoren la latencia, restricciones de ancho de banda de terceros, pérdida de paquetes y configuraciones generales para uso del enlace de datos.

El segundo mecanismo de optimización es una funcionalidad que interactúa o mejora el modo del comportamiento de la propia capa de aplicación, también se conoce a este mecanismo como aceleración de la aplicación. La mayoría de los aceleradores proporcionan una combinación de optimización WAN y aceleración de aplicaciones, así como la distribución de contenidos, que puede considerarse una forma de aceleración de aplicaciones. Las capacidades de aceleración de aplicaciones interactúan en el contexto de aplicaciones y protocolos de aplicaciones específicos para mejorar el rendimiento de los usuarios que acceden a aplicaciones y contenido a través de una WAN, incluyendo comportamientos específicos del protocolo, como transacciones directas o multiplexadas.

La aceleración de la aplicación (específica de aplicaciones) y la optimización de la WAN (agnóstica de las aplicaciones) son dos capas separadas de optimización que comúnmente coexisten en el mismo dispositivo acelerador y se aprovechan mutuamente. Cuando se combinan, la aceleración de la aplicación y la optimización de la WAN pueden superar una serie de desafíos que afectan el rendimiento de las aplicaciones en entornos WAN. [6]

En pocas palabras, los aceleradores hacen que la WAN sea un lugar propicio para el rendimiento de las aplicaciones y que los protocolos de aplicación funcionen mejor sobre las redes de datos. Este capítulo examina cómo los aceleradores hacen que los protocolos de aplicación funcionen mejor en las redes examinando técnicas de optimización que minimizan el impacto del ancho de banda y la latencia, incluidas las técnicas de aceleración reactiva y proactiva, como la de caché, la optimización de lectura anticipada y la precarga, predicción de mensajes, canalización, etc. Las optimizaciones discutidas en este capítulo sirven como base para la aceleración de aplicaciones específicas. Estas técnicas se combinan con los principios de optimización WAN que se revisaron brevemente en los capítulos anteriores para formar la base de las soluciones de aceleración, que ayudan a asegurar

que las redes tengan la capacidad necesaria para entregar un alto rendimiento a las aplicaciones centralizadas, datos y contenidos en un entorno infraestructura global. [6][7]

6.1. DESCRIPCIÓN DE LA ACELERACIÓN ESPECÍFICA DE APLICACIONES

La aceleración específica de aplicaciones se refiere a los mecanismos empleados en la capa de aplicación dentro de un acelerador para mitigar la latencia de la aplicación, mejorar los tiempos de transferencia de datos y proporcionar servicio fuera de banda (*offline*) en caso de interrupción de la red. El nivel de aceleración específica de la aplicación aplicada a una aplicación específica depende directamente de la aplicación, de los protocolos mismos y de lo que se puede hacer dentro de los límites de la corrección, coherencia e integridad de los datos. [6]

Algunos protocolos de aplicación son robustos y complicados, requiriendo de un alto grado de intercambio de mensajes entre nodos. Tales protocolos pueden ser ricos en características y robustos pero producen un rendimiento bajo en la WAN cuando no se usa alguna forma de aceleración. En muchos casos, las aplicaciones que dependen de estos protocolos necesitan que se despliegue la infraestructura necesaria para soportar esas aplicaciones en ubicaciones locales en las oficinas remotas para proporcionar niveles adecuados de rendimiento. La implementación de una infraestructura distribuida puede superar las limitaciones de rendimiento de las aplicaciones que hacen uso de la WAN, pero también conduce a inversiones de capital y gastos de administración de TI considerablemente más altos. [6][7]

La aceleración de la aplicación puede clasificarse en dos principales tipos: la aceleración reactiva y la aceleración proactiva. La aceleración reactiva se refiere al comportamiento empleado por un dispositivo acelerador cuando se observa un tipo específico de información. Estos son comportamientos y funciones que se implementan al encontrar una petición específica de un usuario determinado. La aceleración proactiva, por otra parte, se refiere al comportamiento empleado por un dispositivo acelerador basado en la configuración del administrador. Con una aceleración proactiva, los aceleradores pueden pre-configurarse con instrucciones sobre cómo manejar determinados tipos de datos, mensajes o peticiones específicos o protocolos específicos, y sobre qué conjuntos de datos se distribuirán de forma proactiva en toda la estructura de los aceleradores. Ambos enfoques trabajan de forma coherente y en conjunción con la optimización de la WAN para garantizar mejoras significativas en el rendimiento de los usuarios que acceden a archivos, contenido y aplicaciones a través de una WAN. [6]

6.2. ALMACENAMIENTO EN CACHÉ ESPECÍFICO PARA APLICACIONES

El almacenamiento en caché es un método comúnmente utilizado en los sistemas informáticos y de red para ayudar a compensar las solicitudes de I/O (Escritura y Lectura) que tienen un rendimiento más lento y limitado al dispositivo más lento en el trayecto de la información. El almacenamiento en caché se refiere a la introducción de un buffer intermediario entre dos dispositivos dentro de una ruta de procesamiento, donde el buffer intermedio puede retener una copia de los datos encontrados en el dispositivo más lento, aunque a una capacidad mucho menor. [6]

Por ejemplo, el almacenamiento en caché se emplea entre la CPU de un ordenador personal o servidor y la memoria principal, que actúa como almacenamiento primario. Aunque comúnmente consideramos que la memoria principal es de alto rendimiento, la memoria principal es dramáticamente más lenta que la CPU misma. Al introducir un caché entre la CPU y la memoria principal, el contenido extraído de la memoria se puede almacenar de forma directa en la memoria caché y recuperarse desde allí por la CPU (en lugar de pagar la penalización de rendimiento de buscar los datos directamente desde la memoria). Esto ayuda a mejorar drásticamente el rendimiento de las aplicaciones que se ejecutan en un PC o servidor. De manera similar, la caché se usa comúnmente en otras áreas de un sistema, incluyendo el subsistema de disco, que proporciona niveles mucho mayores de rendimiento y mejores tiempos de respuesta en comparación con la búsqueda continua de datos desde las unidades de disco mecánicas. [6][7]

De forma similar a cómo se implementa el almacenamiento en caché en un PC o en un servidor, el almacenamiento en caché específico para aplicaciones se puede implementar como una función dentro de un acelerador en la red para permitir que el acelerador mantenga copias locales de objetos previamente solicitados (reactivos) al acelerador WAN por medio de una configuración administrativa dentro del equipo (proactiva). Esto minimiza el número de solicitudes de objetos que requieren transmisión a través de la red de baja velocidad, además de proporcionar mejoras en el rendimiento. El almacenamiento en caché para las aplicación también ayuda a descargar otros aceleradores en la ruta de red al servidor de origen y proporciona una cantidad significativa de reducción de la carga de trabajo en el propio servidor de origen, ya que no tiene que servir copias redundantes de datos.

6.3. VENTAJAS DEL ALMACENAMIENTO EN CACHÉ ESPECÍFICO PARA APLICACIONES

El proceso de almacenamiento en caché de objetos que se acceden por un protocolo de destino se considera un proceso específico de la aplicación. Para la aceleración de aplicaciones mediante el almacenamiento en caché, los protocolos como HTTP, HTTPS y el *Common Internet file System* (CIFS), o incluso los protocolos de transmisión de medios como el Protocolo de transmisión en tiempo real (RTSP), se convierten en el objetivo de procesamiento primario del caché. Una caché específica para aplicaciones almacena no sólo los objetos interceptados, sino también cualquier información de *metadata* y de directorios relacionados asociados con la ubicación del objeto. Al mantener un repositorio de objetos previamente accedidos y la *metadata* asociada, el acelerador está en posición de determinar no sólo la validez de los objetos en caché, sino también tener la seguridad de cuando emplear optimizaciones para las conexiones en curso y futuras. [6][7]

El caché para aplicaciones dentro de un acelerador ofrece más que sólo copias de objetos en caché que atraviesan sus interfaces de red. Este proceso se lo describe de la siguiente manera:

a) En primer lugar, se debe ejecutar un proceso de almacenamiento en caché de un contenido específico durante la solicitud inicial coloca el contenido en la memoria y disco del acelerador.
b) En segundo lugar, el contenido debe almacenarse en un lugar que sea rápidamente accesible en caso de que el acelerador reciba una petición posterior.
c) Por último, en el caso de una petición de contenido que existe en el acelerador, el acelerador debe ser capaz de validar la frescura del objeto dado antes de servir el objeto al usuario solicitante.

Esto se hace para asegurar que los datos que hayan cambiado o tengan atributos diferentes no se entreguen incorrectamente a los usuarios. Cada uno de estos pasos permite que la caché de aplicaciones interactúe con el usuario solicitante, el servidor de origen y la aplicación con el objetivo de acelerar las solicitudes de objeto sobre el protocolo de destino. [6][7]

La caché de aplicaciones es útil para cualquier usuario cuya solicitud atraviesa un protocolo de destino. La capacidad del acelerador para interceptar los mensajes del protocolo intercambiados entre un usuario solicitante y el servidor de origen se inspecciona y compara con las solicitudes que pueden haber atravesado previamente el acelerador. Una vez validado, cada objeto almacenado

en el optimizador es entregado a la LAN de una manera más eficiente y rápida, ya que se ha evitado enviar la solicitud a través de la WAN, mitigando así la penalización de rendimiento de la WAN para la mayoría de las transacciones. Aunque el usuario podría no saber que su solicitud fue interceptada y procesada por un acelerador, su solicitud de información se ha atendido de forma más rápida que lo que lo hubiera hecho el servidor de origen remoto a través de la WAN. La capacidad de hablar el protocolo de la aplicación permite a los administradores de aplicaciones y de red implementar el caché de aplicaciones con poco o ningún cambio en el servidor o cliente de origen. [6]

Las cuatro entidades que más se benefician de una caché de aplicaciones son el usuario, la WAN, los aceleradores intermedios y el servidor de origen. El usuario experimenta una notable mejora en el rendimiento al solicitar objetos que son optimizados por un acelerador, porque la mayoría de la carga de trabajo se realiza localmente en la LAN a velocidades cercanas a la de la LAN.

Con el almacenamiento en caché de aplicaciones, se necesitan menos datos para recorrer la WAN. Esto se traduce directamente en ahorros de ancho de banda en la WAN y también minimiza la cantidad de tráfico que debe manejar cualquier acelerador intermedio que se despliegue en la red entre el acelerador más cercano al usuario y al servidor de origen. De forma similar, el servidor de origen probablemente verá una disminución sustancial en el número de solicitudes que debe prestar servicio, lo que permite mayores niveles de escalabilidad con la infraestructura existente en el servidor de origen. De esta manera, los aceleradores que proporcionan aceleración de aplicaciones y caché de objetos como un componente de una solución que mejora el rendimiento no sólo mejoran el rendimiento del usuario, sino que también descargan dramáticamente la red y el servidor de origen. [6][7]

6.4. VALIDACIÓN DEL CACHÉ Y FRESCURA DEL CONTENIDO

Un acelerador que proporciona almacenamiento en caché para aplicaciones valida los elementos de contenido almacenados en sus discos en el momento en que recibe una solicitud que coincide con las propiedades del elemento que está almacenado en la caché. La validación de contenido en caché impide que el contenido obsoleto se sirva a cualquier usuario solicitante y garantiza la entrega de información actualizada y coherente desde una perspectiva de protocolo de aplicación. La validación de contenido es una característica que se requiere que el acelerador realice al almacenamiento en caché de aplicaciones y en los siguientes párrafos se demuestra por qué es necesario.

Para demostrar el impacto de la publicación de contenido obsoleto para un usuario solicitante, considere el impacto en el negocio de una hoja de cálculo de nómina desactualizada que se está prestando a un contador que solicita la hoja de cálculo desde el servidor de nómina. La hoja de cálculo, que está alojada en el servidor de origen, puede tener una "frescura" de minutos u horas, sin embargo si un caché de aplicación no comprueba la frescura de la hoja de cálculo, cualquier cambio de último minuto que no exista en la hoja de cálculo obsoleta podría hacer que un empleado reciba un cheque de pago incorrecto. Otro ejemplo de la importancia de la validación de caché, supongamos que un empleado de la corporación comprueba su estado de cuenta bancaria en línea a través de un navegador web. Sin saberlo, el empleado puede estar accediendo a través de un servicio de caché de contenido y recibir información del estado perteneciente a otro empleado de la corporación, o por el contrario, la información que se acaba de presentar desde la caché de la aplicación puede haber sido almacenada en la caché el día anterior, lo que lleva al empleado a creer que su cuenta tiene un valor diferente al que realmente está allí.

Cada aplicación determina cómo el acelerador que proporciona caché para aplicaciones debe validar los elementos de contenido que se han almacenado en caché, para mantener la coherencia y la corrección. Cada método de validación implica la interacción del protocolo entre el cliente y el servidor de origen, en este escenario el acelerador es capaz de interceptar e inspeccionar la información con el propósito de validar objetos almacenados en la caché basado en la *metadata* de la información almacenada y la respuestas del servidor de origen. Los aceleradores que realizan el almacenamiento en caché para aplicaciones deben adherirse estrictamente a la estructura del protocolo y de las aplicaciones, utilizando los mensajes de protocolo y de aplicación existentes como medio para validar la frescura de los objetos almacenados en caché. Aunque el comportamiento de un acelerador que tiene caché para aplicaciones podría ser diferente de un protocolo a otro, o de aplicación a aplicación (cada uno tiene sus propios requisitos y métodos de validación), el proceso conceptual de validación de objeto en caché sigue siendo el mismo. Cada petición interceptada por el acelerador será analizada y asociada con cualquier contenido en caché local coincidente, es decir se compara con el contenido proporcionado como parte de la respuesta del servidor de origen. [5][6][7]

Cada método de validación de información dentro de un acelerador con caché para aplicaciones debe tener la capacidad de comunicarse nativamente con el servidor de origen según sea necesario e intercambiar información específica de la aplicación y del contenido que requiere validación. Para protocolos tales como HTTP y HTTPS, una caché de aplicaciones web genérica inspecciona la

respuesta del servidor de origen a la petición proxy de un cliente, comparando el texto dentro del encabezado con la información de encabezado registrada de la solicitud anterior que colocó el contenido en el almacenamiento de la caché. Para el tráfico de servidor de archivos CIFS, que no tiene métricas de frescura asociadas al protocolo, se debe comprobar primero con el servidor de origen si el objeto ha cambiado en función de marcas de tiempo, tamaño u otros mecanismos de control de versiones. [6][7]

Los dos niveles de control de frescura implementados para las solicitudes son los siguientes:

- **Coherencia estricta de contenido:** El objeto debe ser validada directamente contra el servidor de origen por el acelerador.
- **Coherencia estricta de contenido:** El objeto puede ser validada sobre la base de las respuestas se envían al usuario por el servidor de origen.

El protocolo CIFS, por ejemplo, necesita del uso de una estricta comprobación de coherencia porque no incluye de forma nativa información dentro de los intercambios de mensajes que define la coherencia del contenido al que se accede, además con CIFS, los usuarios de muchas ubicaciones potenciales tienen la capacidad de aplicar funciones de lectura y escritura al contenido almacenado en el servidor de origen. Mediante CIFS, debe suponerse que el objeto se podría haber editado desde el acceso anterior CIFS, por lo que el acelerador debe estar preparado para ajustar dinámicamente el nivel de caché y optimización aplicada a los de objetos basándose en lo que se puede hacer de manera segura y sin violar los requisitos de corrección del protocolo. El acelerador también debe tomar posesión de la validación de contenido y aprovechar esta información como parte de su modelo de optimización adaptable. [5][6]

Los protocolos como HTTP y HTTPS suelen emplear coherencia no estricta y validación de caché porque la información sobre la validez del contenido se suele llevar en mensajes de protocolo que se intercambian entre clientes y servidores. HTTP y HTTPS se utilizan comúnmente en solicitudes unidireccionales, lo que significa que el contenido que se sirve generalmente no se edita en el borde de la red. Los protocolos tales como HTTP y HTTPS pueden basarse en etiquetas de contenido dentro de los mensajes intercambiados entre los nodos para determinar la frescura de la información almacenada en el caché para las aplicaciones.

Algunas funciones de almacenamiento en caché de protocolo en el optimizador pueden configurarse para evitar completamente la función de validación de frescura de contenido del caché,

pero este tipo de configuración expone al cliente al riesgo de acceder al contenido obsoleto desde la caché de aplicaciones. Este tipo de configuración es común con HTTP e Internet, donde el ahorro de ancho de banda es más importante que la frescura de la información.

La Figura 3 ilustra una coherencia no estricta en la que un acelerador que emplea el almacenamiento en caché para aplicaciones es capaz de examinar los intercambios de mensajes del cliente y del servidor para comprender los períodos de validación del contenido. Esta figura muestra un ejemplo de un golpe de caché (parte superior del ejemplo) y un error de caché (parte inferior del ejemplo). Cualquier información intercambiada entre el usuario y el servidor de origen que atraviesa el acelerador es un candidato para la optimización a través de compresión, optimizaciones de flujo y otras técnicas de optimización de WAN. [6]

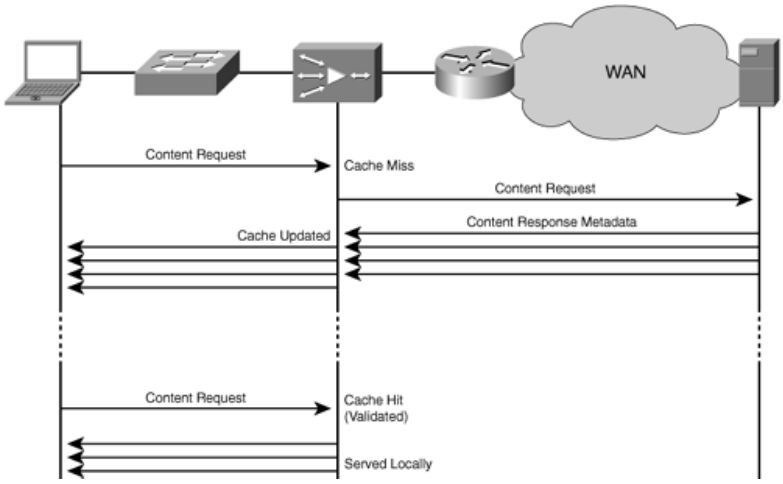


Figura 3: *Validación de Cache con Coherencia No estricta*
Fuente: Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.

La Figura 4 ilustra el modelo de coherencia estricta, donde el empleo de un acelerador de almacenamiento en caché de aplicaciones no es capaz de examinar los intercambios de mensajes de cliente y servidor para comprender los períodos de validación de contenido. Para estas aplicaciones y protocolos, el acelerador debe validar el contenido contra el servidor de origen utilizando otros medios tales como marcas de tiempo u métodos de control de versiones. Esta figura muestra un ejemplo de error de caché (parte superior del ejemplo) y un error de caché (parte inferior del ejemplo). Cualquier información intercambiada entre el usuario y el servidor de origen que atraviesa el acelerador es un candidato para la optimización a través de compresión, optimizaciones de flujo y otras técnicas de optimización de WAN. [6][7]

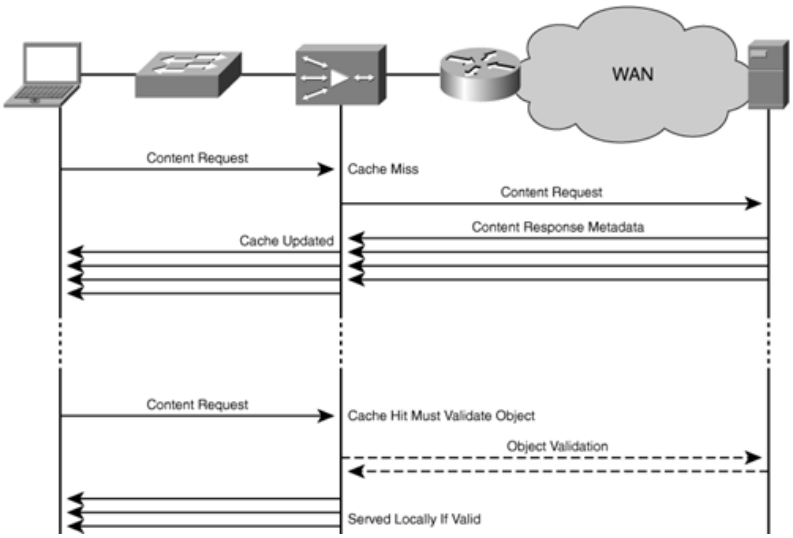


Figura 4: *Validación de Cache con Coherencia estricta*
Fuente: Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.

Las siguientes secciones examinarán el almacenamiento en caché y la optimización con más detalle en relación con CIFS y HTTP.

6.4.1. CIFS

El contenido relacionado con archivos que atraviesa un enlace WAN sobre CIFS puede variar desde unos pocos bytes hasta varios gigabytes de tamaño. El protocolo CIFS se utiliza más comúnmente entre los clientes y servidores basados en Microsoft Windows, proporciona un protocolo para acceder a recursos de red y recursos compartidos de archivos. Dado que CIFS trabaja para hacer accesibles partes de un sistema de archivos local a través de una red, la estructura del sistema de archivos debe ser soportada a través de la red entre el cliente y el servidor. Como tal, el protocolo CIFS es bastante robusto, proporcionando más de 100 diferentes tipos de mensajes que pueden ser intercambiados y distribuidos en diez o más dialectos diferentes (versiones del protocolo). [5][6]

Un acelerador u optimizador que proporciona aceleración y almacenamiento en caché para CIFS debe ser consciente de cómo los clientes y los servidores negocian el dialecto que se utiliza y cómo manejar los diferentes tipos de mensajes que se intercambian entre cliente y servidor. Algunos mensajes pueden ser suprimidos de forma segura o manipulada localmente por el acelerador, sin comprometer la integridad y coherencia de los datos. Otros mensajes son tan críticos para la integridad y coherencia de los datos que deben ser manejados de forma síncrona por el servidor de origen y el acelerador debe estar preparado para optimizar la entrega de dichos mensajes al servidor. [6]

Cuando un acelerador con capacidades de almacenamiento en caché y aceleración CIFS está en la ruta de la sesión del cliente a servidor, se aplican varios beneficios de reducción de tráfico WAN y mejoras de rendimiento, entre los que se incluyen:

- El número de mensajes intercambiados entre el cliente y el servidor se reducirá considerablemente, ya que el acelerador puede minimizar con seguridad la cantidad de conversación que debe intercambiarse a través de la WAN. Esto ayuda a mejorar el tiempo de respuesta al cliente, minimiza la cantidad de trabajo que debe realizar el servidor de origen y anula la penalización de latencia WAN en el rendimiento del usuario.
- El contenido del servidor al que accede el cliente puede ser capturado y almacenado en caché por la caché de aplicaciones si es seguro hacerlo. Esto permite al acelerador mitigar la necesidad de transferir archivos a través de la WAN a los que se ha accedido anteriormente, asumiendo que el usuario está autenticado y autorizado y que el objeto en caché se valida con éxito contra el servidor de origen.

El aspecto más importante de una caché es su capacidad para almacenar objetos enteros o porciones de objetos que son utilizables en una capa de aplicación. Esta copia almacenada se convierte en la principal fuente de beneficios al implementar una caché específica para aplicaciones. El contenido no necesita ser servido por el servidor de origen una segunda vez si se cumplen los requisitos de validación y frescura, lo que también se traduce en una reducción del número de bytes que deben ser enviados a través de la WAN, lo que potencialmente baja el ancho de banda y la latencia. Los beneficios obtenidos de un caché específico de aplicaciones aumenta con cada solicitud adicional recibida para una parte de contenido que ya reside dentro de la memoria caché y puede ser validada como coherente en comparación con la copia en el servidor de origen. Los beneficios de la capacidad de la memoria caché para volver a servir cualquier pieza determinada de contenido almacenado en la caché puede rastrearse de forma lineal, en función del número de solicitudes que se procesan después del almacenamiento inicial del contenido. [6][7]

Además de la reducción del número de solicitudes que deben atravesar la WAN, la sobrecarga de mensajería CIFS también se reduce significativamente por la caché de aplicaciones y la optimización del protocolo. Es posible que algunos mensajes no necesiten recorrer la WAN para que un cliente pueda acceder con éxito a un contenido determinado o buscar en un directorio, mientras que otros requieren intercambios de mensajes con el servidor de origen. Por ejemplo, un acelerador puede haber realizado una tarea de lectura anticipada y almacenar temporalmente los resultados de los

datos para su uso posterior por un usuario. Una solicitud para estos datos se puede satisfacer localmente, eliminando así la necesidad de enviar la petición de lectura del usuario al servidor de origen.

Muchos aceleradores que optimizan CIFS pre-almacenan grandes cantidades de información sobre la estructura de directorios y colocan estos datos en el borde de la red. Si un usuario navega por una estructura de directorios, los datos de recorrido de directorio pueden ser servidos desde el acelerador si se consideran válidos. La mayoría de los aceleradores que optimizan el recorrido de directorio usan una ventana muy corta de validación en datos de recorrido de directorio (por ejemplo, 30 segundos) para permitir que las actualizaciones de directorio sean recibidas por el usuario de manera oportuna.

Son aquellos mensajes que no necesitan comunicarse con el servidor de origen, los que la caché de aplicaciones interpretará y responderá localmente desde el acelerador. Sin embargo, los mensajes que son críticos para la integridad, seguridad o estado de los datos deben atravesar la WAN para asegurar la integridad del objeto CIFS y la estructura del protocolo. Tales mensajes incluyen aquellos que proporcionan negociación de protocolo, autenticación de usuario, autorización de usuario, bloqueo de archivos, solicitudes de archivo abierto y operaciones de escritura. Sin embargo, estos mensajes pueden optimizarse a través de técnicas de optimización de WAN que operan de una manera agnóstica de la aplicación, incluyendo compresión, optimización de flujo y mitigación de pérdidas.

CIFS crea de forma nativa una tremenda cantidad de tráfico adicional en intercambios de cliente a servidor. En muchos casos, una operación aparentemente simple como un archivo abierto de un documento de 1 MB puede requerir que se intercambien más de 500 mensajes entre el cliente y el servidor. En un entorno WAN, el rendimiento máximo que se puede lograr es afectado directamente por la cantidad de latencia total de la red y se calcula fácilmente como el número de mensajes multiplicado por la latencia de operación.

Por ejemplo, en una WAN de 100 ms de ida y vuelta, 500 mensajes que se intercambian pueden llevar a más de 100 segundos de latencia para abrir un documento simple. Con la aceleración CIFS y el almacenamiento en caché CIFS, los clientes se benefician de la eliminación o reducción de ciertos tipos de tráfico adicional y potencialmente de la eliminación de la transferencia redundante del archivos solicitado. Reducir la cantidad de latencia experimentada por el cliente durante los

intercambios de cliente a servidor mejorará el rendimiento general percibido de una aplicación determinada.

Un caché específico de aplicaciones para CIFS dentro de un acelerador normalmente interceptará peticiones sobre los puertos TCP estándar usados por CIFS-TCP / 139 y TCP / 445 y muchos pueden ser configurados para acelerar CIFS en puertos no estándar si es necesario, lo cual no es común. El objetivo principal de la memoria caché para aplicaciones CIFS es minimizar el consumo de ancho de banda y mejorar los tiempos de respuesta para el contenido solicitado o las transferencias de archivos. Cuanto mayores sean los objetos almacenados en la memoria caché, más rápido serán los tiempos de respuesta para todos los clientes que soliciten información que esté previamente almacenada en la caché. Basado en esto, el almacenamiento en caché del acelerador CIFS es útil en entornos en los que se desea la consolidación de servidores de archivos, ya que el rendimiento proporcionado por el acelerador es similar al proporcionada por un servidor de archivos local.

Los administradores de sistemas suelen utilizar CIFS para distribuir grandes objetos a través de la red. Algunas de las aplicaciones más comunes que aprovechan grandes objetos de contenido incluyen:

- Aplicaciones de diseño
- Instaladores de aplicaciones
- Videos e Imagen de alta calidad
- Información Cartográfica, etc.

Los objetos que crean estas aplicaciones pueden ser tan pequeños como unos pocos cientos de kilobytes o tan grandes como varios cientos de megabytes o gigabytes. Las aplicaciones que se centran en los gráficos y las imágenes implican una gran cantidad de detalle y, por lo tanto consumen una gran cantidad de capacidad de almacenamiento. Con el aumento de los requisitos de resolución de imagen, la cantidad de uso del disco consumido aumenta también. Las aplicaciones médicas que involucran imágenes consumen grandes cantidades de almacenamiento en disco, dependiendo del tipo de imagen que se realiza. Los administradores de redes empresariales también aprovechan habitualmente CIFS para distribuir actualizaciones de software, parches del sistema operativo, o respaldos completos de los servidores.

Para CIFS, el acelerador intercepta y examina cada mensaje que se transfiere entre el cliente y el servidor. El acelerador determina si el cliente y el servidor están realizando un intercambio de mensajes y cuál es la transacción que se está realizando, lo que permite al acelerador tomar

decisiones inteligentes acerca de cómo procesar y optimizar la conversación. Mientras observa el tráfico entre hosts, el acelerador puede eliminar con seguridad algunos mensajes después de determinar que no son requeridos por el servidor de origen.

Cada archivo que atraviesa el acelerador puede o no ser almacenado en discos locales del acelerador en función del nivel de optimización que se desee aplicar. Para el contenido que se almacena en los discos del acelerador, el contenido podría ser objetos de contenido enteros o partes de una determinada pieza de contenido. La caché específica de aplicaciones puede realizar de forma proactiva la lectura anticipada de la solicitud del cliente para recuperar el contenido que falta en la memoria caché. En general, este método dual de contenido mejorará el rendimiento para el usuario y rellena dinámicamente las partes restantes de los datos que faltan del archivo parcial en caché. [5][6]

Una caché específica para aplicaciones puede almacenar en caché sólo partes de un archivo determinado en base a las necesidades de la aplicación solicitante; un archivo parcialmente en caché no demuestra una caché de aplicación defectuosa. En la Figura 5 se ilustra un objeto parcialmente en caché que se sirve de la caché del acelerador, mientras tanto el acelerador recupera los segmentos que faltan del contenido del servidor de origen.

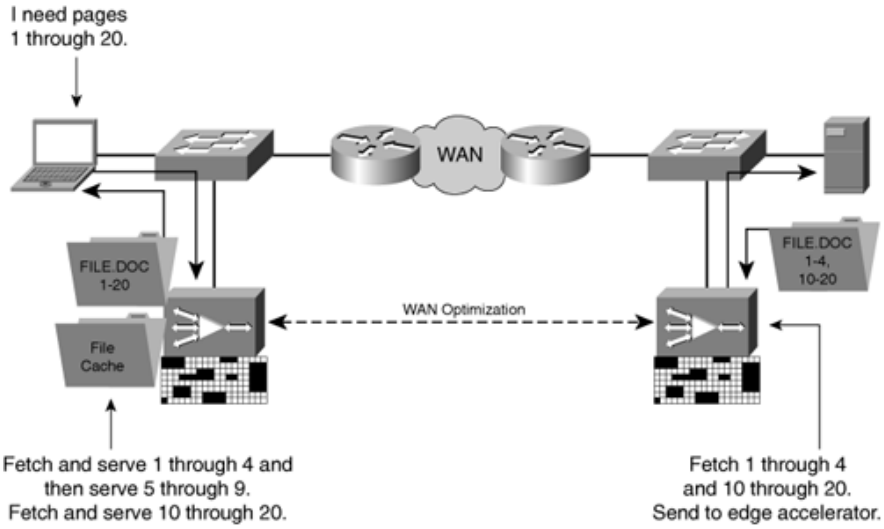


Figura 5: *Servidor de caché para objetos de Lógica Parcial*
Fuente: *Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

El almacenamiento en caché y la funcionalidad de lectura anticipada se combinan para proporcionar una mejora sustancial de rendimiento al acceder a objetos través de la WAN. Para las solicitudes de contenido que pueden ser servidos a nivel local, el acelerador puede responder a solicitudes de

datos cuando sea seguro hacerlo. Para las solicitudes de contenido que no pueden ser satisfechas por el acelerador, el acelerador puede optar por aplicar optimizaciones adicionales, tales como la pre-lectura, que ayuda a satisfacer las futuras solicitudes de las secciones faltantes del objeto solicitado. [6]

6.4.2. HTTP

Las funcionalidades de caché de aplicaciones para protocolo HTTP permite el almacenamiento en caché de forma reactiva de los objetos que se solicitan bajo demandad. Para realizar la intercepción del tráfico HTTP a través del almacenamiento en caché tradicional, el acelerador debe tener la capacidad de hablar de forma nativa el protocolo HTTP. Aunque HTTP y HTTPS comparten los mismos métodos de transferencia fundamentales, HTTPS se trata de manera diferente por los aceleradores. El cifrado de la sesión es la principal diferencia entre los protocolos, y algunos aceleradores puede convertirse en un *man in the middle* que un proceso de descifrado, optimización y re-codificación. La respuesta a las solicitudes que sirven localmente a la red se vuelven a cifrar, y cualquier petición que se deba enviar hacia la WAN se vuelve a cifrar para mantener la estructura de seguridad. [6][7]

La caché HTTP puede interceptar transparente o explícitamente las peticiones HTTP de dispositivos tales como ordenadores personales o incluso asistentes digitales personales. Una petición HTTP consiste en una conexión establecida comúnmente a través del puerto TCP 80, desde un cliente a un servidor de origen dado, la mayoría del tráfico web de las organizaciones utiliza el puerto 80 y es comúnmente clasificado como tráfico de intranet. Si el servidor de origen reside en la Internet pública, este tráfico se considera como tráfico de Internet. Las solicitudes HTTP no se limitan a puerto TCP 80, y es común observar conexiones a los puertos TCP 8000, 8001, 8080, 3128, y muchos otros puertos. Cada uno de estos puertos y cualquier otro número de puerto personalizado puede ser interceptado por el acelerador y almacenado en el caché de aplicaciones HTTP.

Dependiendo de la versión de HTTP soportado entre el cliente y el servidor web de origen, las solicitudes pueden ser secuenciales, al igual que con la versión 1.0 de HTTP. HTTP 1.0 también requiere que se establezca una nueva conexión TCP entre el cliente y el servidor para cada objeto que se está solicitando. HTTP 1.1 supera estas limitaciones y proporciona soporte para múltiples conexiones simultáneas que se establezcan entre el cliente y el servidor, reduciendo así al mínimo la cantidad de tiempo que se gasta establecimiento de conexiones TCP. [6][7]

Las peticiones secuenciales de HTTP 1.0 requieren de una respuesta para cada solicitud. Antes de realizar la próxima petición al servidor web la petición anterior debe ser respondida y después se puede establecer una nueva conexión TCP. Las peticiones HTTP 1.0 son muy ineficientes, debido a la naturaleza de inicio y parada del protocolo y del inicio y parada de TCP como protocolo de transporte. Cuando se combina con una WAN que tiene una latencia significativamente alta, HTTP 1.0 es extremadamente ineficiente al cliente solicitante y puede convertirse en un cuello de botella.

Cuando una caché específica de aplicaciones HTTP se combina con otras técnicas de optimización, tales como lectura anticipada y compresión, el tráfico total que atraviesa la WAN entre un acelerador de *core* y acelerador de borde se minimiza significativamente. Aunque muchos servidores web hoy en día se aplican protocolos de compresión como gzip, la memoria de caché todavía basa sus decisiones en las propiedades de cabecera proporcionadas por el servidor web. Muchas veces, los objetos servidos por el servidor web contienen patrones de datos que el acelerador puede aprovechar a través de algoritmos de supresión de los datos. Un objeto entregado desde un servidor web puede contener patrones que los aceleradores de *core* y borde ya han registrado, eliminando de esa manera los patrones de tráfico que se encuentran dentro de la respuesta. Cuando se combinan estas técnicas, el almacenamiento en caché de HTTP, lectura anticipada, la optimización del transporte, compresión, y la supresión de los datos redundante, se proporciona una gran mejora a la experiencia web del usuario final. [6][7]

6.4.3. MEDIA STREAMING: RTSP, HTTP, Y FLASH

El *streaming* de contenidos se ha convertido en uno de los métodos más populares de la comunicación corporativa. El *streaming* de contenidos lleva el presentador a su audiencia final. Los siguientes cuatro proveedores fueron fundamentales para el éxito del *streaming* de información:

- Microsoft Corp., con su servidor de Windows Media y basada en el cliente de Windows Media Player.
- RealNetworks, Inc., con su servidor Helix y su cliente basado en RealPlayer.
- Apple, Inc., con su servidor Darwin Stream y QuickTime Player del lado del cliente.
- Adobe Systems, Inc., con su Flash Media Server y su cliente de Flash Player. [6]

Cada uno de estos proveedores de *streaming media* requiere dos componentes comunes: un servidor que proporciona el contenido sobre el protocolo de transporte nativo del vendedor y un reproductor de medios dedicado o *plug-in* instalado navegador del ordenador del cliente para

decodificar y reproducir el contenido de manera efectiva. Desde la perspectiva de la aplicación, el *streaming* tiene la capacidad de entregar a las empresas comunicaciones a un nuevo nivel.

El *streaming* de contenidos se utiliza dentro de redes empresarial con menos frecuencia que el vídeo bajo demanda. El *streaming* de contenidos en vivo implica mucho más planificación y sincronización de agendas. El acelerador se convierte comúnmente en un punto de división del *streaming* en vivo en el centro de datos y en el borde de la red. El proceso de división del *streaming* permite que el acelerador use una sola señal en vivo desde el código fuente del servidor de los medios y disponer de varios flujos de información a los clientes que acceden al contenido desde el otro lado del acelerador. Este método de división de flujos de contenidos permite que el acelerador sirva a un gran número de clientes sin requerir múltiples copias de la misma información a través de la WAN. La caché de aplicación enfocada en *streaming*, no obtiene una copia del evento en vivo a medida que atraviesa el acelerador sino que facilita la capacidad de distribuir una sola alimentación a múltiples usuarios simultáneamente. En la Figura 6 se muestra cómo un pequeño número de flujos de un servidor de origen se puede utilizar para entregar contenido en vivo a un gran número de usuarios. [6]

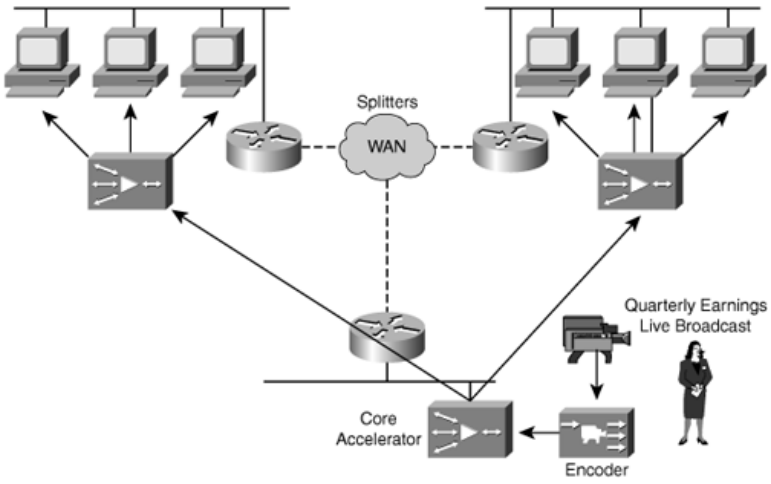


Figura 6: *División del streaming para minimizar el consumo de Ancho de Banda WAN*
Fuente: *Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

Cuando un acelerador que es compatible con optimizaciones para RealNetworks, Windows Media y QuickTime recibe la solicitud inicial RTSP para *streaming* de contenido, busca información del reproductor de medios del cliente dentro de las primeras transacciones de la solicitud. Una vez que

el acelerador ha identificado el reproductor de medios del cliente, dirige la petición al motor de *streaming* correspondiente dentro del acelerador.

El contenido tipo *video-on-demand* que ha sido pre-posicionado en el acelerador a través de protocolos como HTTP o CIFS, también aprovecha los componentes de optimización WAN. Aunque el contenido de vídeo por naturaleza ya está comprimido, los patrones de datos repetidos se pueden eliminar de manera segura a través de la supresión de los datos cuando atraviesa la WAN.

Al igual que con otros métodos de comunicación que acceden a través de la WAN, el contenido tipo *video-on-demand* está sujeto a los mismos riesgos de información obsoleta y falta de actualización. Si se accede al contenido de vídeo bajo demanda a lo largo de CIFS o HTTP, entonces el acelerador confirma que el contenido que se sirve es la última versión disponible desde el servidor de origen. Si el contenido se va a acceder en un protocolo nativo tal como RTSP, entonces el acelerador de aplicaciones primero valida el factor de frescura del contenido con el servidor de origen antes de su entrega al cliente solicitante. [6][7]

Para los eventos que se transmiten en vivo, no se aplican las preocupaciones de la actualización del contenido; el evento en vivo se entrega a los clientes que participan en el momento exacto que el contenido está siendo transmitido en toda la red. El acelerador puede servir simplemente como una plataforma para la entrega optimizada a través de la división de flujos para reducir al mínimo el consumo de ancho de banda en la WAN para un gran número de espectadores remotos.

6.4.4. APLICACIONES DE BASES DE DATOS BASADAS EN WEB

Tres de las aplicaciones Web más comúnmente usadas en las empresas hoy en día incluyen Bases de Datos Oracle y Siebel, así como SAP AG, de la alemana SAP. Cada una de estas aplicaciones tiene las siguientes características:

- La capacidad de permitir el acceso del cliente a una aplicación a través de un navegador web como una alternativa o conjuntamente con la interacción de una aplicación de escritorio que está preinstalada o que se instala en el proceso.
- Un modelo de acceso HTTP con autentificación, con la opción de hacer la transición a HTTPS para mayor seguridad.
- La utilización de *applets* de Java para acceder a las aplicaciones mediante el navegador web del cliente.

- Tiempos de espera potencialmente largos para el acceso a las aplicaciones debido a *applets* de Java que deben ser transferidos desde el servidor de base de datos hacia el navegador web del cliente a través de la WAN

Las aplicaciones basadas en web pueden ser abordados por los aceleradores a través de dos modelos diferentes. La primera implica el almacenamiento en caché de la aplicación descrita anteriormente. Los aceleradores que realizan el almacenamiento en caché de la aplicación son capaces de mantener una copia en caché de objetos, como *applets* de Java, que se han transferido con anterioridad. Esto permite que el acelerador reduzca al mínimo la transferencia redundante de objetos a través de la red. El segundo enfoque implica el uso de los componentes de optimización WAN a través de los aceleradores, lo que incluyen la optimización del transporte de protocolo (superar la pérdida de paquetes, latencia y rendimiento), además de las técnicas de compresión, y la supresión de datos. En esta sección se ilustra ambos modelos, mostrando los beneficios que cada enfoque proporciona a estas aplicaciones críticas para el negocio. [6]

Oracle, Siebel y SAP soportan el uso del protocolo HTTP para proporcionar acceso distribuido simplificado a través de una red corporativa. Las peticiones HTTP que son atendidas por el servidor de base de datos al cliente solicitante llevan varios objetos que pueden ser almacenados en la caché. Dependiendo de cómo esté configurado el servidor de base de datos, el *applet* de Java que se entrega primero al cliente será un archivo JAR, en concreto el Oracle Jinitiator. Para los clientes Siebel se entrega *applets* de Java y los objetos gráficos, mientras que SAP utiliza objetos JS, GIF y CSS. Dependiendo de la configuración de la aplicación y del servidor de origen, los objetos que sirvieron para un cliente pueden o no ser identificados como contenidos almacenables en caché.

Cuando se utiliza la aceleración HTTP y el almacenamiento en caché de un acelerador, tres sencillos pasos ayudan a identificar cómo el contenido puede ser almacenado en caché:

1. Habilitar el acelerador para interceptar las solicitudes el servidor de base de datos. Esto podría implicar cambios de configuración especiales sobre el acelerador, permitiendo la interceptación y aplicación de aceleración de tráfico en puertos no estándar tal como el puerto 8802. Algunos aceleradores reconocerán automáticamente los puertos que se utilizan comúnmente por los fabricantes de esas aplicaciones.
2. Habilitar el registro de transacciones en el acelerador. Los registros de transacciones ayudan a identificar fácilmente qué tipos de contenidos pueden ser almacenados en caché y qué tipos no lo son.

3. Tener acceso a un analizador de captura de paquetes o protocolo como *Ethereal* o *Wireshark*, para inspeccionar las respuestas HTTP de las cabeceras de objetos que atraviesan la red entre el servidor de base de datos y el cliente.

Una vez que haya configurado el acelerador con la aceleración HTTP, incluyendo el almacenamiento en caché para interceptar el tráfico de la base de datos del cliente, se comprueba las estadísticas del acelerador para determinar el ahorro de caché inicial y los ahorros por compresión. Hay la posibilidad de que no se requiera cambios de configuración adicionales en el acelerador, lo que resulta en una implementación más simple. Aunque es poco probable, puede encontrarse con que el servidor no tiene parámetros de la cabecera del mensaje que sean susceptibles de ser almacenados en la caché del acelerador y sea imposible atender peticiones actuales y futuras desde el acelerador.

Cuando se inspeccionando los registros de transacciones reunidos por el acelerador, el cliente genera tráfico de prueba que debe solicitar el mismo *applets* de bases de datos al menos dos veces. Si el usuario ha solicitado un objeto dado sólo una vez, no hay manera de validar si la caché es capaz de almacenar el objeto para futuras peticiones. Los registros de transacciones ayudan a identificar rápidamente el nombre del archivo y la URL asociada a la solicitud, así como la forma en que el acelerador debe procesar el contenido requerido durante la segunda solicitud. Los registros de transacciones que señalan que la segunda petición fue distinta de la primera requieren de una investigación por parte del administrador de la red. [5][6]

Muchos de los objetos generados para páginas dinámicas proceden de direcciones URL estáticas del servidor de base de datos. Los aceleradores que implementan estándares de almacenamiento en caché para HTTP deben generar cambios en el servidor o balanceador de carga del centro de datos, para mejorar el almacenamiento en caché. Si no se realizan estos cambios, otros componentes tales como la supresión de datos seguirán proporcionando potencialmente enormes niveles de optimización y ahorro de ancho de banda.

La Figura 7 ilustra las capacidades de almacenamiento tradicional en caché web inteligente de los aceleradores con el servidor de base de datos en un centro de datos central y los clientes que solicitan información desde sucursales remotas.

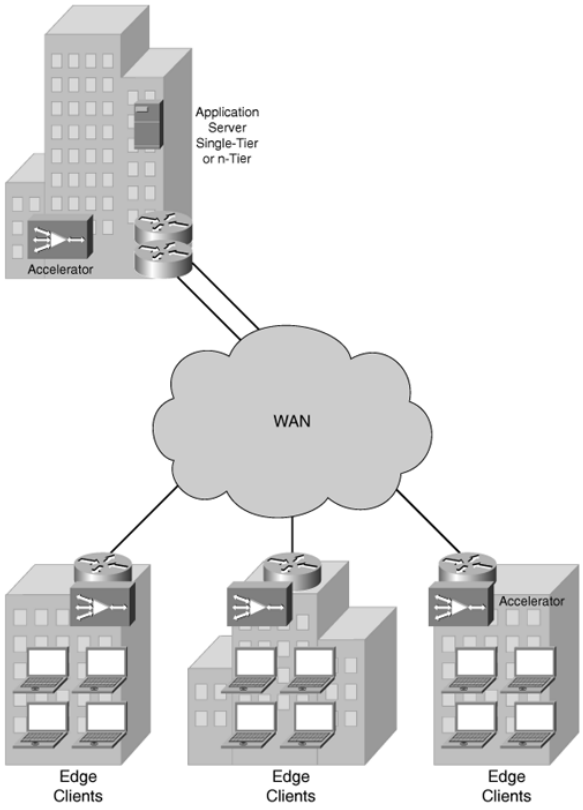


Figura 7: *Despliegue de aceleradores para aplicaciones Web y de Base de Datos*
Fuente: *Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

Muchos de los datos que se entregan por una aplicación de base de datos se obtienen de ubicaciones estáticas. Estos objetos no cambian, y generalmente son los *applets* de Java requeridos por el cliente para ejecutar la aplicación en el navegador del cliente. Estas solicitudes de objetos recurrentes generan una caché de aplicación que funcionan muy bien. Una vez que el acelerador de *core* o central ha observado e identificado un objeto que se genera desde el servidor de aplicaciones, puede solicitar un proceso de caché en los aceleradores de borde. El contenido dinámico que entrega el servidor de base de datos tiene un componente de información y *applets* de Java, del cual el mayor componentes en tamaño de información son los *applets* que necesita el navegador para presentar la información en el navegador del cliente. El desempeño que observa el cliente en un ambiente con aceleradores puede resultar de hasta diez veces más rápido que una conexión no optimizada mediante aceleradores WAN. [6][7]

La Figura 8 ilustra una aplicación de base de datos tradicional con aceleradores de núcleo y de borde que proporcionan almacenamiento en caché específica para aplicaciones.

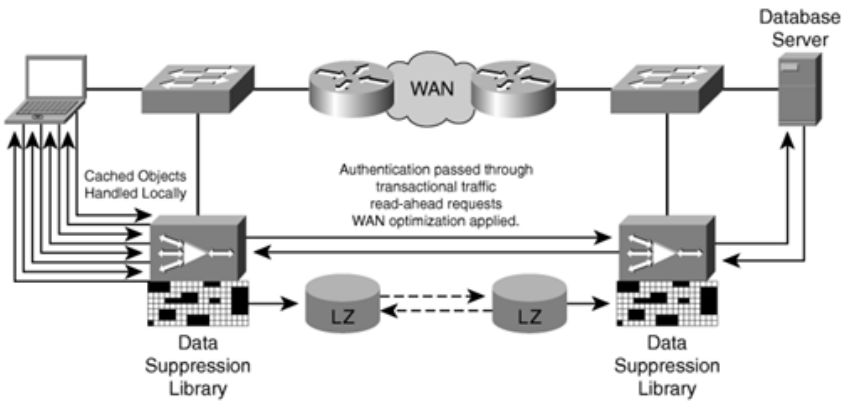


***Figura 8**: Aceleración específica para aplicaciones de Base de Datos*
***Fuente:** Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

La optimización de aplicaciones para empresas se realiza a través de una jerarquía de componentes, que incluye el almacenamiento en caché de aplicaciones, la aceleración de protocolo, y los componentes de optimización WAN. Cuando se combinan, estas funciones proporcionan un ahorro de ancho de banda, reducción de latencia y mejora del rendimiento general.

7. LIMITACIONES DE LOS ENLACES DE DATOS WAN.

En los capítulos anteriores se ha discutido cómo alinear los recursos de red con las prioridades del negocio y los requerimientos de las aplicaciones, así como las técnicas que se pueden aplicar dentro de la red y los dispositivos aceleradores para mejorar el rendimiento general de una aplicación a través de la red. Las técnicas empleadas en la red, como la calidad de servicio (QoS), pueden ayudar a alinear la capacidad de la red con los requisitos de rendimiento de las aplicaciones o ajustar el encolamiento y la programación para una aplicación específica que puede ser sensible a la latencia. El almacenamiento en caché para datos de aplicaciones, lectura anticipada, predicción de información y otras técnicas pueden ayudar a mitigar la utilización innecesaria de ancho de banda para transferencias de objetos redundantes y también minimizan la latencia al manejar la mayor parte de la carga de trabajo localmente o de una forma más optimizada. [5][7]

Sin embargo, se puede emplear capas más genéricas de optimización, que pueden funcionar simultáneamente en varias aplicaciones. Este tipo de optimización, comúnmente llamado optimización WAN, se refieren generalmente a las funciones que se encuentran comúnmente en dispositivos aceleradores (tales como equipos tipo *appliance* o módulos integrados a los *routers*) estos mecanismos superan las limitaciones de rendimiento causadas por los protocolos de transporte, pérdida de paquetes, y limitaciones de capacidad. [6]

Las capacidades de optimización de la WAN hacen que la WAN sea un lugar más tolerable para las aplicaciones, eliminando las barreras de rendimiento que crea la WAN. Por ejemplo, la compresión de red avanzada se puede aplicar para mejorar el rendimiento al minimizar la cantidad de datos que necesitan recorrer la WAN. Un beneficio secundario de esto es que se necesitan menos intercambios de datos sobre la WAN, con lo que se mitiga la latencia asociada con el número de intercambios de ida y vuelta que hubieran sido necesarios. Por otro lado, la optimización TCP se utiliza comúnmente para permitir a los nodos utilizar de forma más eficiente los recursos disponibles y minimizar el impacto por la pérdida y latencia en la WAN.

Este capítulo examina cómo las capacidades de optimización de la WAN superan las barreras de rendimiento creadas por las condiciones de la WAN. Se debe tener en cuenta que, en términos de productos de aceleración, se puede utilizar las capacidades de optimización WAN en conjunto con otras tecnologías de optimización que son específicas para la aplicación, como se describe en el capítulo 6 de este trabajo. Por otra parte, suponiendo que la arquitectura del acelerador es

transparente, se puede utilizar estas tecnologías de optimización junto con funciones orientadas a la red que proporcionan visibilidad y control.

7.1. LIMITACIONES DEL TRANSPORTE DE PROTOCOLO

La mayoría de la gente se pregunta cómo TCP (u otros protocolos de transporte) podría convertirse en un cuello de botella, simplemente porque da la ilusión de que siempre funciona. En una red de trabajo tipo intranet, debe existir una capa entre las aplicaciones y la infraestructura de red. Esta capa, llamada *capa de transporte,* no sólo ayuda a asegurar que los datos se mueven entre los nodos, sino que también ayuda a entender a los nodos como la red está funcionando de modo que puedan adaptarse a la misma.

Aunque la capa de transporte es un candidato poco probable para problemas de rendimiento en las aplicaciones, puede convertirse en uno principalmente porque los protocolos de transporte actualmente en uso fueron diseñados en 1981 [6]. La demanda de las aplicaciones y topologías de red actuales difieren mucho de las redes de principios de los años ochenta. Por ejemplo, 300 baudios se consideraban una velocidad rápida en el momento en que se creó TCP. La congestión se aceleró en gran parte al cambio hacia redes complejas, de alta velocidad y jerárquicas como Internet, que está plagada de sobre-escritura, agregación y el uso simultáneo de millones de usuarios que se enfrentan por el ancho de banda disponible. Las aplicaciones en 1981 eran comúnmente aplicaciones orientadas al texto (y en gran parte orientadas a terminales), mientras que hoy en día incluso el usuario corporativo más elemental puede mover fácilmente archivos desde decenas hasta cientos de megabytes de tamaño durante una sola transferencia.

Aunque la red ha cambiado, TCP sigue siendo relevante en el entorno de una red dinámica y cambiante. TCP ha sufrido sólo cambios menores en los últimos 25 años, y esos cambios son en forma de extensiones en lugar de reescrituras mayores del protocolo. Aunque hay algunos protocolos de transporte más modernos que tienen raíces en TCP, muchos se consideran proyectos de desarrollo solamente y actualmente tienen un despliegue limitado en el mercado. [6]

Otra consideración importante en relación con los actuales entornos de red y aplicaciones empresariales es el costo y la capacidad disponible de la tecnología WAN frente a la disminución de los costos de la tecnología LAN. En efecto, la capacidad de ancho de banda de red ha aumentado constantemente durante los últimos 20 años; sin embargo, el costo de la capacidad de ancho de

banda de la LAN ha disminuido a una velocidad mucho más acelerada que el costo del ancho de banda de la WAN.

La disparidad cada vez mayor entre el ancho de banda WAN y LAN presenta varios desafíos, especialmente relacionados con el rendimiento de las mismas. Las aplicaciones y el acceso al contenido se han vuelto más fácil para los usuarios de la LAN, ya que el ancho de banda de LAN ha aumentado, sin embargo, el mismo nivel de acceso a aplicaciones y contenido no se ha vuelto más accesible para los usuarios de la WAN, debido a la diferencia entre el costo y la capacidad de ancho de banda que la WAN ha experimentado. En pocas palabras, la tasa de aumento de ancho de banda que se encuentra en la WAN no está a la par con el ritmo de la LAN, y esto crea desafíos de rendimiento para los usuarios que se ven obligados a acceder a la información a través de la WAN. De esta manera, la LAN permite un acceso más rápido a los datos y de una forma más intensiva. Al mismo tiempo, la WAN no ha crecido en el mismo grado desde una perspectiva de capacidad. [6][7]

Las técnicas de optimización de la WAN se consideran complementarias a técnicas como la optimización de red (*QoS* y enrutamiento optimizado) y aceleración de aplicaciones (almacenamiento en caché y otras optimizaciones como lectura anticipada). Por ejemplo, una caché de aplicaciones de capa de objetos puede aprovechar las tecnologías de compresión durante la distribución de contenido para mejorar el rendimiento y asegurar que el historial de compresión se rellene con el contenido relevante si la transferencia de los objetos en cuestión tiene lugar a través de la WAN.

En la dirección contraria, los usuarios que tienen un tipo de relación de lectura-escritura con un objeto que se ha abierto a través de la caché de objetos de un acelerador, donde dicho objeto ha sido validado contra el servidor de origen (en el caso de un archivo en caché), pueden aprovechar el historial de compresión y las optimizaciones de protocolo (como la optimización *write-back*) para mejorar el rendimiento de escritura y ahorrar ancho de banda.

Las técnicas de compresión pueden aprovecharse para minimizar el ancho de banda consumido y eliminar patrones de bytes repetitivos que previamente se hayan identificado. Esto no sólo ayuda a garantizar que el ancho de banda WAN se conserve, sino que también sirve para mejorar el rendimiento de la experiencia de usuario porque se necesita mucho menos ancho de banda. En consecuencia, se deben producir menos intercambios de paquetes antes de que se complete la operación.

La optimización de la WAN ayuda a superar las limitaciones propias del uso de este tipo de conexiones, al mismo tiempo que mantiene los costos de la WAN, preserva la inversión y proporciona una solución para consolidar el servidor distribuido, el almacenamiento y la infraestructura de aplicaciones. La optimización y compresión de los protocolos de transporte (es decir, la optimización de la WAN) garantizan que los recursos se utilicen de manera eficaz y eficiente mientras se superan las barreras de rendimiento en la capa de transmisión de datos [7]. La aceleración de aplicaciones funciona para eludir las barreras de rendimiento de la capa de aplicación. Estas tecnologías son todas independientes pero se pueden combinar coherentemente para formar una solución, como se muestra en la Figura 9.

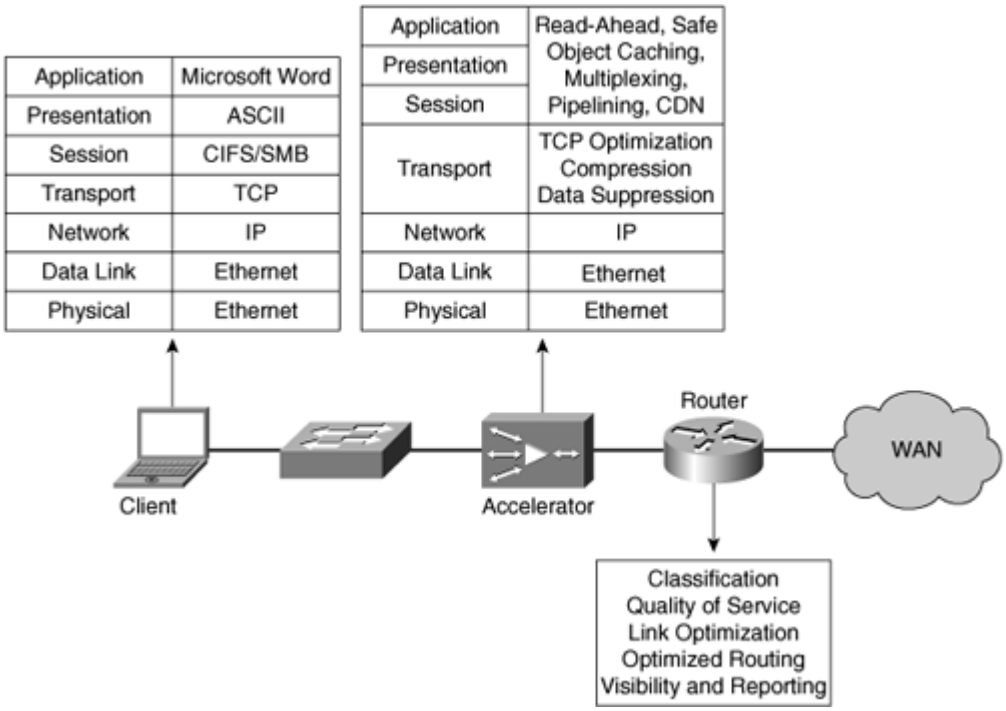


Figura 9: *Aceleración de aplicaciones y jerarquía de optimización de WAN*
Fuente: *Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

7.2. FUNDAMENTOS DEL PROTOCOLO DE CONTROL DE TRANSMISIÓN

TCP es el protocolo de transporte más utilizado para aplicaciones que se ejecutan en redes empresariales e Internet en la actualidad. TCP proporciona las siguientes funcionalidades:

- Servicio orientado a la conexión entre los procesos de aplicación de dos nodos que están intercambiando datos.
- Entrega garantizada de datos entre estos dos procesos.
- Detectar el de ancho de banda para evitar la congestión y utilizar con equidad el ancho de banda disponible basado en la utilización y la capacidad de la WAN.

Antes de que se puedan enviar datos entre dos procesos de aplicaciones diferentes entre dos nodos dispares, primero se debe establecer una conexión. Una vez establecida la conexión, TCP proporciona una entrega fiable y garantizada de datos entre los dos procesos de la aplicación.

7.2.1. SERVICIOS ORIENTADOS A CONEXIÓN

La conexión TCP se establece mediante un acuerdo de *handshake* de tres vías que se produce entre dos *sockets* de dos nodos que desean intercambiar datos. Un *socket* se define como el identificador de red de un nodo junto con el número de puerto que está asociado con el proceso de aplicación que desea comunicarse con un par remoto. El uso de sockets TCP se muestra en Figura 10.

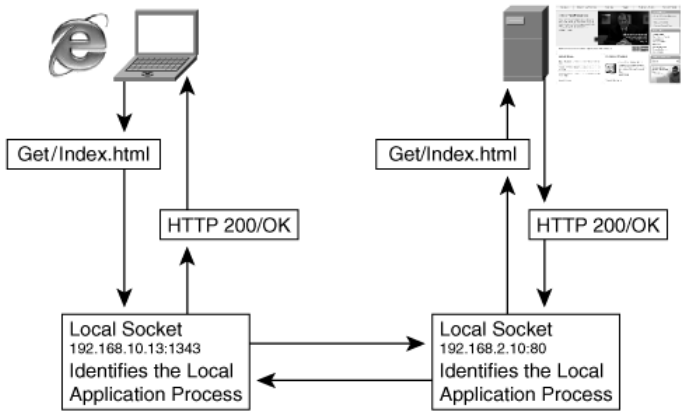


Figura 10: *Socket TCP*

Fuente: *Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

Durante el establecimiento de la conexión, los dos nodos intercambian información relevante para los parámetros de la conversación. Esta información incluye:

Tipo de información	*Descripción*
Origen y destino de los puertos TCP	Los puertos que están asociados con los procesos de aplicación en cada uno de los nodos que deseen intercambiar datos de la aplicación.

Número de secuencia inicial	Cada dispositivo notifica al otro el número de secuencia se debe utilizar para el inicio de la transmisión.
Tamaño de la ventana	Es la cantidad de datos que el nodo receptos puede retener de forma segura en su búfer de recepción de socket.
Opciones	Campos de cabecera opcional que se utilizan comúnmente para extender el comportamiento de TCP. Por ejemplo, este podría incluir características tales como el ajuste de ventanas y el reconocimiento selectivo que no fueron incluidos como parte de la RFC de TCP, pero puede aumentar el comportamiento de TCP (una lista autorizada de opciones de TCP se puede encontrar en http://www.iana.org/assignments/tcp-parameters)

Tabla 2: *Información de intercambio durante la conexión TCP [6]*

Por ejemplo, si un usuario de Internet quiere usar Internet Explorer para acceder a una URL, el equipo del usuario primero tendrá que resolver el nombre de la URL en una dirección IP y, a continuación, intentar establecer la conexión TCP con el servidor web que aloja la URL utilizando el conocido puerto para HTTP (puerto TCP 80) a menos que se especifique un número de puerto distinto. Si el servidor web que aloja la URL está aceptando conexiones en el puerto TCP 80, la conexión se establecerá con éxito. Durante el establecimiento de la conexión, el servidor y el cliente se informan entre sí cuántos datos pueden recibir en su memoria intermedia de socket (tamaño de la ventana) y qué número de secuencia inicial utilizar para la transmisión inicial de datos. A medida que se intercambian los datos, este número se incrementa para permitir al nodo receptor conocer el orden apropiado de los datos. Durante la vida de la conexión, TCP utiliza la funcionalidad de *checksum* para proporcionar integridad a los datos. [5][6]

Una vez establecida la conexión entre los dos nodos (direcciones IP) y dos procesos de aplicaciones (puertos TCP), los procesos de la aplicación que utilizan esos dos puertos en los dos nodos pueden comenzar a intercambiar datos de la capa de aplicación. Por ejemplo, una vez establecida la conexión, el usuario puede enviar una solicitud GET al servidor web al que está conectado para comenzar a descargar objetos desde una página web, o también el usuario puede comenzar a intercambiar mensajes de control utilizando SMTP o POP3 para transmitir o recibir un mensaje de correo electrónico. El establecimiento de la conexión TCP se muestra en la Figura 11.

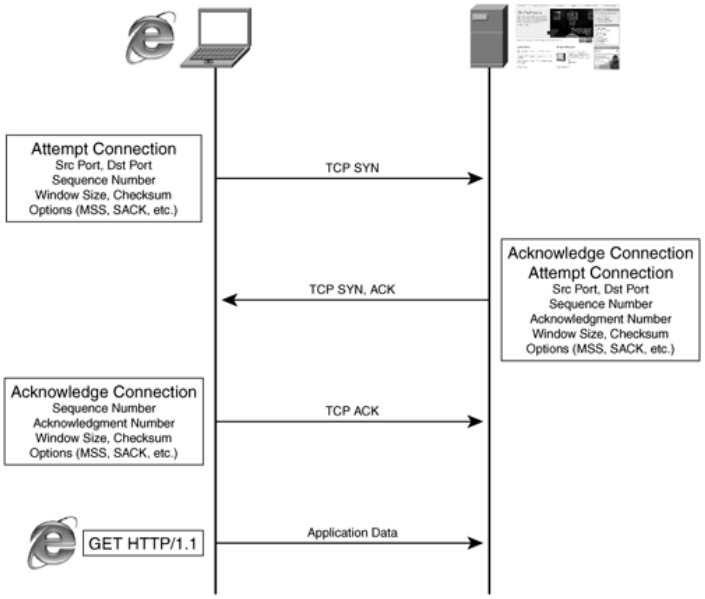


Figura 11: *Establecimiento de la conexión TCP*
Fuente: *Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

7.2.2. GARANTIZANDO LA ENTREGA

Una vez que la transmisión comienza, los datos de la aplicación se liberan del buffer de la aplicación en el proceso de transmisión al buffer del socket del nodo. Entonces el protocolo TCP negocia la transmisión de datos desde el buffer de transmisión del socket al nodo receptor (es decir, la liberación del buffer) basándose en la disponibilidad de recursos en el nodo receptor para recibir los datos según lo dictado por el tamaño de ventana inicialmente anunciado y el tamaño de ventana actual. Dado que los bloques de datos de aplicación pueden ser bastante grandes, TCP realiza la tarea de dividir los datos en segmentos, cada uno con un número de secuencia que identifica el ordenamiento relativo de las porciones de datos que se han transmitido. Si el nodo recibe los segmentos fuera de orden, TCP puede reordenarlos de acuerdo con el número de secuencia. Si los *buffers* TCP se llenan por una de las siguientes razones, podría ocurrir una condición de bloqueo:

- **El buffer de transmisión TCP se llena:** El buffer de transmisión en el nodo se puede llegar a llenar si las condiciones de red impiden la entrega de datos o si el destinatario es saturado y no puede recibir datos adicionales. Aunque el nodo receptor no puede recibir más datos, se puede permitir que las aplicaciones continúen agregando datos al buffer de transmisión para esperar el servicio. Con la información bloqueada esperando en el buffer de transmisión, y siendo la información incapaz de ser transmitida, las aplicaciones en el nodo transmisor pueden quedarse bloqueadas (es decir, pausas momentáneas o prolongadas en la transmisión). En tal situación,

no se pueden escribir nuevos datos en el buffer de transmisión a menos que haya espacio disponible en ese buffer, que generalmente no se puede liberar hasta que el destinatario sea capaz de recibir más datos o la red sea capaz de entregar datos nuevamente. [6][7]

- **El buffer de recepción TCP se llena:** comúnmente es causado porque la aplicación receptora no es capaz de extraer datos del buffer con la suficiente rapidez. Por ejemplo, un servidor sobrecargado, es decir, uno que está recibiendo datos a una velocidad mayor que la velocidad a la que puede procesar datos, presentará esta característica. A medida que el buffer de recepción se llena, no se pueden aceptar nuevos datos de la red para este socket y se deben descartar, lo que indica un evento de congestión para el nodo transmisor. [6][7]

En la Figura 12 se muestra cómo TCP actúa como un tampón intermedio entre la red y las aplicaciones dentro de un nodo.

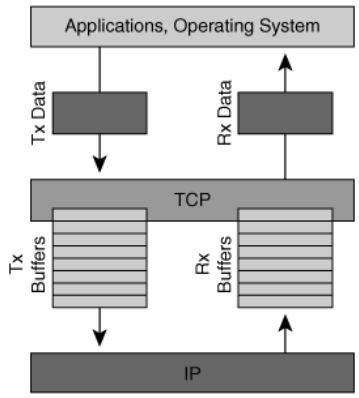


Figura 12: *Buffer TCP entre la red y las aplicaciones*
Fuente: *Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

Cuando los datos se colocan con éxito en un nodo receptor de buffer, TCP genera un acuse de recibo (ACK) con un valor relativo a la cola de la secuencia que acaba de ser recibida. Por ejemplo, si el número de secuencia inicial era "1" y se transmitía 1 KB de datos, cuando los datos se colocan en el buffer de recepción del receptor, la pila TCP del destinatario emitirá un ACK con un valor de 1024. Dado que las aplicaciones hoy en día son generalmente capaces de extraer datos casi inmediatamente de un buffer de recepción TCP, es probable que el reconocimiento y el relieve de la ventana se ejecuten simultáneamente.

El siguiente segmento que se envía desde el nodo transmisor tendrá un número de secuencia igual al número de secuencia anterior más la cantidad de datos enviados en el segmento anterior (1025 en este ejemplo) y sólo se puede transmitir si hay capacidad de ventana disponible en el nodo

receptor según lo dictado por los acuses de recibo enviados desde el destinatario. A medida que se reconocen los datos y se incrementa el valor de la ventana (los datos en la memoria intermedia del socket TCP deben ser extraídos por el proceso de la aplicación, lo que alivia la capacidad del buffer y, por lo tanto, la capacidad de la ventana), el remitente puede enviar datos adicionales al destinatario hasta la capacidad máxima de la ventana del destinatario (el destinatario también tiene la capacidad de enviar actualizaciones de ventanas dinámicas que indican aumentos o disminuciones en el tamaño de la ventana).

Este proceso de transmisión de datos en función de la capacidad del receptor para recibir segmentos previamente reconocidos se conoce comúnmente como la *ventana deslizante TCP*. En esencia, cuando el destinatario continúa recibiendo y reconociendo o notificando de otro modo un aumento en el tamaño de ventana, la ventana en el nodo transmisor cambia para permitir que se envíen nuevos datos. Si en algún punto los *buffers* se llenan o la ventana se agota, el destinatario debe primero dar servicio a los datos que se han recibido previamente y reconocer al remitente antes de que se puedan enviar nuevos datos. Un ejemplo de este proceso se muestra en Figura 13. [6][7]

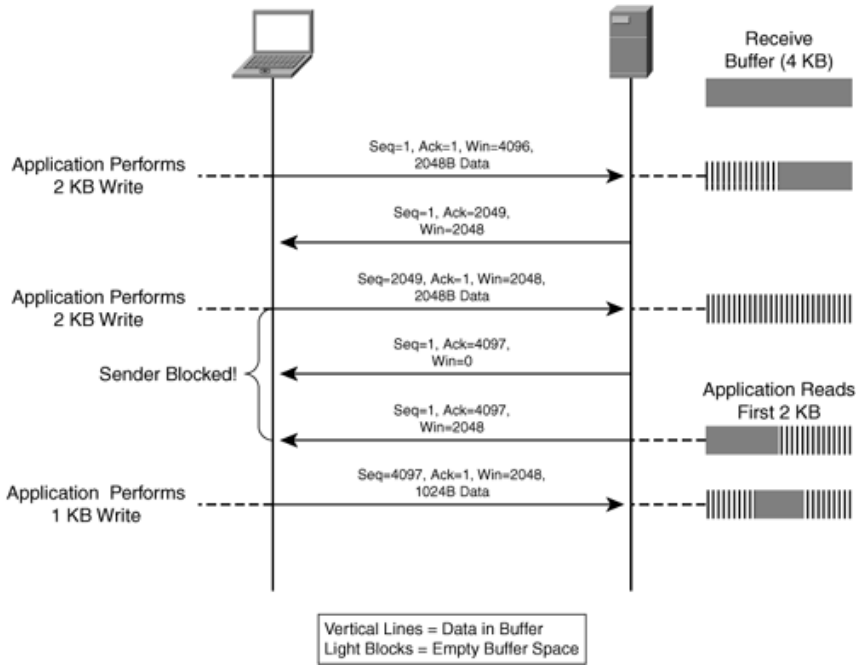


Figura 13: *Operación TCP*

Fuente: Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.

Además, TCP proporciona un indicador que permite que un proceso de aplicación notifique a TCP para enviar datos inmediatamente en lugar de esperar a que una cantidad mayor de datos se acumule en el búfer del socket. Esta bandera, llamada *push,* o PSH, instruye a TCP para enviar de

inmediato todos los datos que están retenidos para un destinatario en particular. Este envío tipo *push* de datos también requiere que se cumplan condiciones previas, incluida la disponibilidad de una ventana en el nodo receptor. Cuando se transmiten los datos, el indicador PSH en el encabezado TCP se establece en un valor de 1, que también indica al destinatario que envíe los datos directamente al proceso de aplicación en lugar de utilizar el buffer de recepción del socket.

Los nodos que transmiten también usan el número de acuse, el número de secuencia y el valor de la ventana como un indicador de cuánto tiempo se conservarán los datos que se han transmitido previamente. Cada segmento que se ha transmitido y está esperando confirmación se coloca en una cola de retransmisión y se considera que no se reconoce por el proceso de solicitud del destinatario. Cuando se coloca un segmento en la cola de retransmisión, se inicia un temporizador que indica cuánto tiempo el remitente esperará un acuse de recibo. Si no se recibe un acuse de recibido, el segmento se retransmite. Dado que el tamaño de ventana es generalmente mayor que un solo segmento, es probable que muchos segmentos estén pendientes en la red esperando confirmación en un momento dado. [6][7]

Desde la perspectiva de la capa de transporte, la pérdida de un segmento puede no impedir que continúe la transmisión. Sin embargo, dado que la capa de aplicación realmente está dictando el comportamiento de la capa de transporte (por ejemplo, un acuse de recibo del protocolo de la capa superior), la pérdida de un segmento puede evitar que la transmisión continúe.

El propósito de esta cola de retransmisión tiene dos aspectos:

- Permitir al nodo transmisor asignar capacidad de memoria para retener los segmentos que se han transmitido previamente. Si se pierde un segmento (congestión, pérdida de paquetes), se puede transmitir desde la cola de retransmisión, y permanecer allí hasta que se reconozca por el proceso de aplicación del destinatario.
- Permitir que el segmento original, una vez colocado en la cola de retransmisión, se elimine de la cola de transmisión original. Esto, en efecto, permite que TCP extraiga continuamente datos del proceso de aplicación de transmisión local sin comprometer la capacidad del nodo transmisor de retransmitir si un segmento se pierde o no se reconoce en el destino.

Un ejemplo de la gestión de retransmisión TCP se muestra en la Figura 14.

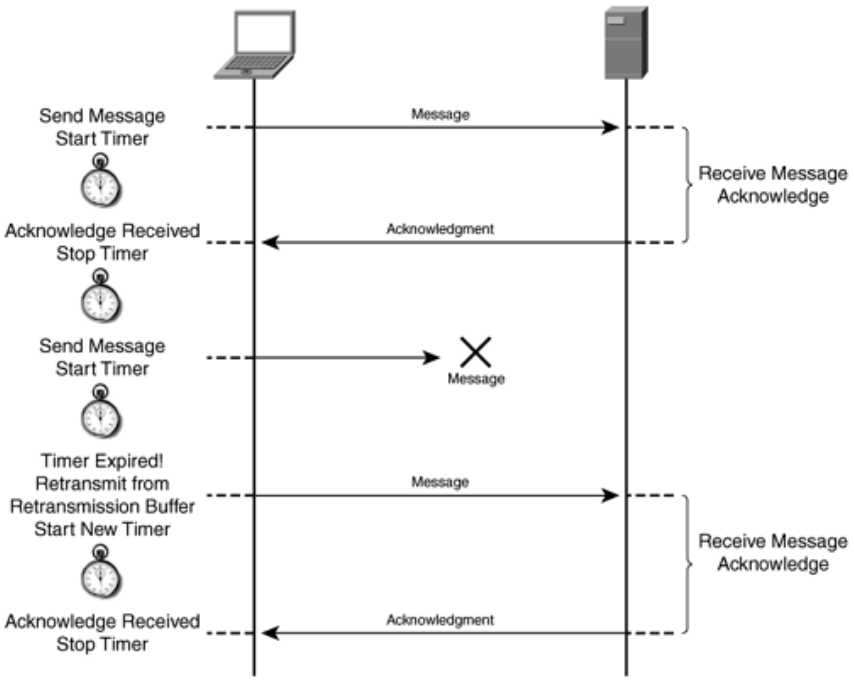


Figura 14: *Gestión de retransmisión TCP*
Fuente: *Application Acceleration and WAN Optimization Fundamentals – Ted Grevers, Jr., Joel Christner.*

7.2.3. DESCUBRIENDO DEL ANCHO DE BANDA

La *ventana deslizante TCP* puede actuar como un mecanismo tipo embudo para asegurar que la transmisión de datos se realiza de tal manera que se alinea con la capacidad de memoria intermedia disponible y la ventana de los dos dispositivos que intercambian datos. También hay mecanismos en TCP que le permiten actuar como un mecanismo de embudo basado en la capacidad de la red y cualquier situación que se encuentre en la red.

En algunos casos, los nodos que intercambian datos pueden enviar más datos de los que la red puede manejar, y en otros casos (que son más comunes hoy en día), los nodos no pueden enviar tantos datos como la red pueda manejar. En el caso de que los nodos puedan transmitir más datos de los que la red está preparada para manejar, se produce congestión y pérdida. TCP tiene mecanismos construidos para detectar estas situaciones. De esta manera, TCP es adaptativo en la medida que cierra la brecha entre los requisitos de transmisión, requisitos de recepción, la congestión, y pérdida de información, tanto para el proceso de aplicación como el de red. Este mecanismo de embudo también proporciona un equilibrio entre las aplicaciones y la red, trabajando para aprovechar continuamente la capacidad de la red disponible mientras intenta maximizar el rendimiento de la aplicación.

Los términos congestión y pérdida se utilizan por separado aquí, aunque la congestión y la pérdida van generalmente de la mano. En algunas situaciones, la congestión puede referirse simplemente a un retraso en el servicio debido a saturaciones de buffer o colas que no están completamente llenas, pero lo suficiente como para retrasar un poco la entrega de un segmento de datos.

TCP proporciona capacidades para responder a las condiciones de la red, permitiéndole así realizar las siguientes funciones básicas pero críticas:

- Encontrar inicialmente un nivel seguro en el que los datos pueden ser transmitidos y adaptarse continuamente a las condiciones de la red.
- Responder a la congestión o pérdida de paquetes a través de la retransmisión y hacer ajustes a los niveles de rendimiento.
- Proporcionar equidad cuando varios usuarios simultáneos compiten por el mismo recurso compartido (ancho de banda).

Estas capacidades se implementan en dos funciones TCP que se discuten en las siguientes dos secciones: inicio lento de TCP y mecanismos para evitar la congestión TCP.

7.2.4. INICIO LENTO DE TCP

TCP es inicialmente el responsable de encontrar la cantidad de capacidad de red disponible para la conexión. Esta función es un mecanismo que se encuentra en TCP y se conoce como inicio lento (*slow start,* también conocido como *el descubrimiento de ancho de banda)* y se emplea en conexiones de vida más larga cuando la ventana disponible cae por debajo de un valor conocido como el *umbral de inicio lento.* [6][7]

El inicio lento utiliza un aumento exponencial en el número de segmentos que se pueden enviar por trayecto de ida y vuelta, este mecanismo se emplea al comienzo de una conexión para encontrar la capacidad de red inicial disponible. En un trayecto de ida y vuelta con éxito, los datos se transmiten basándose en el valor actual de la ventana de congestión (*Congestion Window*s, *cwnd*) y se recibe un acuse de recibo del destinatario. (El *cwnd* se correlaciona con el número de segmentos que se pueden enviar y no se reconocen en la red y es un valor dinámico que no puede exceder el tamaño de ventana del destinatario). El valor de la ventaba de congestión está enlazado a un umbral superior definido por el tamaño de la ventana del receptor y la capacidad de memoria intermedia de transmisión del emisor.

Con el inicio lento de TCP, el nodo transmisor empieza enviando un único segmento y espera confirmación. Tras el acuse de recibo, el nodo transmisor duplica el número de segmentos que se envían y espera el acuse de recibo. Este proceso se produce hasta que se encuentra uno de los dos escenarios siguientes:

- No se recibe un acuse de recibo, lo que indica la pérdida de paquetes o la congestión excesiva
- El número de segmentos que se pueden enviar es igual al tamaño de la ventana del destinatario o igual a la capacidad máxima del remitente

El primer caso sólo se encuentra en las siguientes circunstancias:

- La capacidad de la red es menor que las capacidades de transmisión del remitente y las capacidades de recepción del receptor
- Se detecta un evento de pérdida (ningún acuse de recibo recibido dentro del tiempo asignado al segmento transmitido cuando se coloca en la cola de retransmisión)
- La congestión retrasa la entrega del segmento lo suficiente permitiendo que el temporizador de la cola de retransmisión para el segmento transmitido expire.

El segundo caso se encuentra sólo cuando las capacidades de la red son iguales (sin pérdida o congestión) o mayor que las capacidades de transmisión del emisor y las capacidades de recepción del receptor. En este caso, el remitente o el receptor no pueden capitalizar completamente la capacidad de la red disponible en función de la capacidad de la ventana o de los tamaños del buffer.

El resultado del inicio lento de TCP es un aumento exponencial en el rendimiento hasta la capacidad de red disponible o hasta el rendimiento de transmisión/recepción disponible de los dos nodos dictados por el tamaño del buffer o el tamaño de ventana esperado. Este es un proceso relativamente preciso para encontrar el ancho de banda inicialmente disponible, y generalmente está sesgado sólo en el caso de redes que presenta altas características de pérdida de paquetes, lo que puede cortar el proceso de inicio lento.

7.2.5. MECANISMOS PARA EVITAR LA CONGESTIÓN TCP

Una vez que la conexión TCP sale del proceso de inicio lento (o descubrimiento de ancho de banda), entra en un modo que trata de evitar la congestión *(congestion avoidance)* [6][7]. El *congestion avoidance* es un mecanismo que permite a TCP reaccionar ante situaciones encontradas en la red

como la pérdida de paquetes y la congestión de la señal de retardo en la red, lo que podría ser indicativo de una serie de factores:

- **Cambio de asignación del ancho de banda:** Por ejemplo, un cambio de ancho de banda en una conexión puede resultar en que la red sea capaz de dar un mejor desempeño basado en la dirección y la naturaleza del cambio.
- **Sobresuscripción de la red:** Cuando una conexión de red compartida entre los dispositivos de subida a la red es utilizado por múltiples usuarios simultáneos, puede congestionarse hasta el punto de producir pérdidas o retraso.
- **Congestión de las colas de los dispositivos:** es similar a la sobresuscripción de la red, un dispositivo compartido, como un *router* puede tener sus colas completas hasta el punto de no ser capaz de aceptar nuevos paquetes. Esto también podría equipararse a una configuración de *QoS* que dicta el uso máximo de ancho de banda de una clase de tráfico específica o política de caída para ese tráfico cuando se produce la congestión.
- **Sobrecarga en el destino:** el buffer del socket de destino puede llegar a saturarse debido a la incapacidad de una aplicación para liberar los datos en el momento oportuno, debido potencialmente a que el servidor esta sobrecargado.

Estos son solo algunas de las causas del por qué los paquetes se podrían perder o retrasarse. La buena noticia es que TCP fue diseñado para trabajar en una red que no es confiable y con pérdidas y utiliza estos eventos a su ventaja para ajustar adecuadamente las características de transmisión y adaptarse a las condiciones de la red.

8. ANÁLISIS DE LAS HERRAMIENTAS DE OPTIMIZACIÓN DISPONIBLES EN EL MERCADO

Según el estudio realizado por Joel Snyder para la revista web: http://www.networkworld.com se usaron diferentes equipos de optimización en un escenario laboratorio para evaluar sus capacidades en ocho áreas: rendimiento, gestión del tráfico, visibilidad, gestión de enlace de datos, y facilidad de uso. [8]

El trabajo de investigación diseñó una pequeña red basada en los routers Cisco y firewalls de Juniper para simular el funcionamiento de una red de área amplia – WAN. Para reproducir las condiciones de una red intercontinental, se utilizaron simuladores de enlace para introducir selectivamente los cuellos de botella del ancho de banda, latencia y errores en la red. El objetivo era simular una red de alrededor de 100 sitios conectados a través de túneles IPSec estándar utilizando aproximadamente 45 Mbps de ancho de banda en la WAN. En los bordes de la red, se utilizó el software de virtualización de VMware para albergar diversas herramientas de pruebas de rendimiento de la red, incluyendo productos de código abierto. Los detalles específicos se encuentran en el respectivo artículo y se analizará concretamente los resultados para tener una base para la selección de la herramienta en las pruebas de laboratorio de este trabajo.

8.1. ENTORNO DEL PRUEBAS DE LAS HERRAMIENTAS DE OPTIMIZACIÓN

A medida que las aplicaciones se mueven a la nube, los administradores de red están experimentando cada vez más la demanda de optimizar y administrar las conexiones WAN. La mayoría de las empresas han migrado a aplicaciones basadas en la Web y utilizan mucho los servicios de Internet para el día a día, haciendo que el rendimiento de la red sea un factor clave para la productividad y la satisfacción del usuario final.

Los principales fabricantes han respondido con una gran cantidad de dispositivos destinados a mejorar el rendimiento de la red. Muchos se centran en una sola función, como la compresión, gestión del ancho de banda o la visibilidad del flujo. Según el estudio realizado por la revista digital Network World (http://www.networkworld.com) se realizó un estudio en el que se invitaron a los principales proveedores de optimización de red, entre los que se pueden resaltar los siguientes:

a) Blue Coat Mach5 editions of the SG300-25 and SG900-10

b) Cisco Wide Area Virtualization Engine WAVE-7541, Cisco 4451-AX ISR and 2900-AX ISR.

c) Citrix CloudBridge 2000

d) Exinda Networks Model 6862 and 10862 running x800-series software

e) Ipanema Technologies ip|engine 1000ax and ip|engine 20ax

f) Riverbed Steelhead CXA-5050 and CXA-555

g) Silver Peak Systems VX-1000 and VX-5000.

En este estudio se puso cada producto en el laboratorio para una ronda intensiva de pruebas en tres áreas clave de la optimización de la red: rendimiento, visibilidad y control (se analizan otras áreas en el estudio pero no se incluyen como parte del análisis de este trabajo). También se examinaron las tendencias del mercado de optimización de la red, evaluando cada producto por su idoneidad empresarial, flexibilidad y facilidad de uso. Según el análisis realizado se resalta que la optimización de WAN ha evolucionado para abarcar otras características, como la gestión del tráfico y la visibilidad. Aquí, se encontró por ejemplo que la solución de Riverbed puede trabajar en algunos aspectos, especialmente en la gestión del tráfico que es buena, pero no sobresale de las existentes en el mercado. La visibilidad es limitada de una manera que empuja a los administradores de redes a usar las propias herramientas de Riverbed, en lugar de abrirse al creciente mundo de los productos de análisis de flujo basados en estándares abiertos. Las nuevas características de Riverbed Steelhead, como la selección de rutas WAN, no cumplen con el liderazgo tecnológico de los competidores.

En el campo de la innovación existen productos que destacan como Ipanema Technologies ip|engines y Exinda Networks x800-series. Estos dos productos ofrecen un enfoque holístico a la optimización de redes que no se ve en Riverbed Steelhead. Estos vendedores están claramente pensando más allá de los aspectos básicos que los productos de optimización entregan y se enfocan en la próxima generación de productos de optimización de red. Al ser los nuevos competidores, sin embargo, se encuentran fallas y agujeros en los productos. Por ejemplo, el sistema de gestión de Exinda es débil, mientras que las funciones de integración de red y gestión de tráfico de Ipanema son demasiado rígidas para funcionar bien con algunas redes.

En el aspecto de rendimiento, uno de los competidores que destaca es Silver Peak, por su desarrollo constante. Empatados en el primer lugar en las pruebas de compresión y deduplicación, la serie VX y NX de Silver Peak simplemente hacen que las cosas vayan más rápido. Sin embargo, Silver Peak está teniendo dificultades en sus soluciones de optimización tipo *data center-to-data center*, y la

serie VX y la serie NX necesitan actualizaciones para manejar los requisitos de las redes con muchas sucursales remotas.

WASS de Cisco es una línea de productos tan amplia que es difícil saber qué se puede probar, y cómo clasificarlo. A medida que Cisco WAAS pasa de dispositivos independientes a software integrado, se adquiere toda la potencia y funcionalidades de IOS, pero a costa de una mayor complejidad. En el caso de integrar soluciones de optimización en oficinas remotas rara vez la opción pesa a favor de Cisco para la optimización de la red.

Otro de los competidores es Mach5 de Blue Coat y CloudBridge 2000 de Citrix Systems. Ambos cubren los conceptos básicos y funcionan según lo anunciado, aunque CloudBridge 2000 probablemente necesita un poco más de trabajo antes de ser desplegado. Sin embargo, Blue Coat no ha podido aprender de sus propios expertos en tecnología de cómo integrar la gestión del tráfico con la compresión. CloudBridge 2000 por otro lado sólo incluye la más mínima de las características necesarias para competir en esta área. Ninguna de las dos compañías es una portadora de estándares sobre cómo impulsar un amplio espectro de funciones de optimización de red en oficinas remotas.

A continuación se presenta un análisis más detallado de los productos, dividido por áreas funcionales:

SCORECARD

Product	Blue Coat	Cisco	Citrix	Exinda	Ipanema	Riverbed	Silver Peak
Performance (25%)	3.5	4.5	3	4	4	4.5	4.5
Traffic Management (25%)	3.5	3.5	3.5	4.5	4.5	4	3.5
Visibility (25%)	3	3	3	4.5	4	4.5	4
Enterprise Features (10%)	3.5	4,5	3.5	3	4.5	4.5	3.5
Flexibility and Ease of Use (10%)	4.5	3.5	4	4.5	4.5	4.5	3.5
IPv6 (5%)	5	2	2	5	2	4.5	4
Total	3.6	3.2	3.2	4.3	4.1	4.4	3.9

SCORING KEY > **5:** Exceptional, **4:** Very Good, **3:** Average, **2:** Below Average, **1:** Consistently Subpar

Figura 15: *ScoreCard de los productos de optimización*
Fuente: *http://www.networkworld.com/article/2363942/lan-wan/riverbed-wins-7-vendor-wan-optimization-test.html*

8.2. PRUEBAS DE RENDIMIENTO

El rendimiento es una de las primeras razones por las que los administradores de red empiezan a buscar productos de optimización de red, generalmente estos productos deben utilizar una combinación de técnicas, incluyendo el almacenamiento en caché (generalmente llamado "deduplicación" para distinguirlo del tipo de almacenamiento en caché que hacen los proxies web) Compresión, optimización del protocolo TCP e IP y optimización específica de aplicaciones como se discutió en el Capítulo 8 de este trabajo.

Cada técnica funciona de diferentes maneras, y, dependiendo de su interacción con las aplicaciones, puede ser más o menos beneficiosa para el ambiente del cliente. Por ejemplo, si se mueve archivos grandes de texto o base de datos, la compresión en línea ahorra mucho ancho de banda. Si mueve un archivo de datos grande que sólo cambia un poco cada día, la deduplicación ayuda. Si sus sistemas operativos están ajustados con pequeños tamaños de ventana TCP o están utilizando algunos tipos de control de congestión, la optimización de TCP/IP ayuda bastante.

En el análisis realizado se analiza el rendimiento desde el punto de vista de la experiencia del usuario final. Para esto se enfocan en cinco tipos de tráfico que son representativos de muchas WAN de la empresa: tráfico web encriptado y no cifrado, correo electrónico, terminal remoto (específicamente Citrix Xen Desktop) y Voz sobre IP. Se hicieron prueba a través de cinco tipos de WAN, utilizando un emulador para variar la latencia y la pérdida, simulando el tráfico que atraviesa enlaces WAN de cobre a diferentes velocidades.

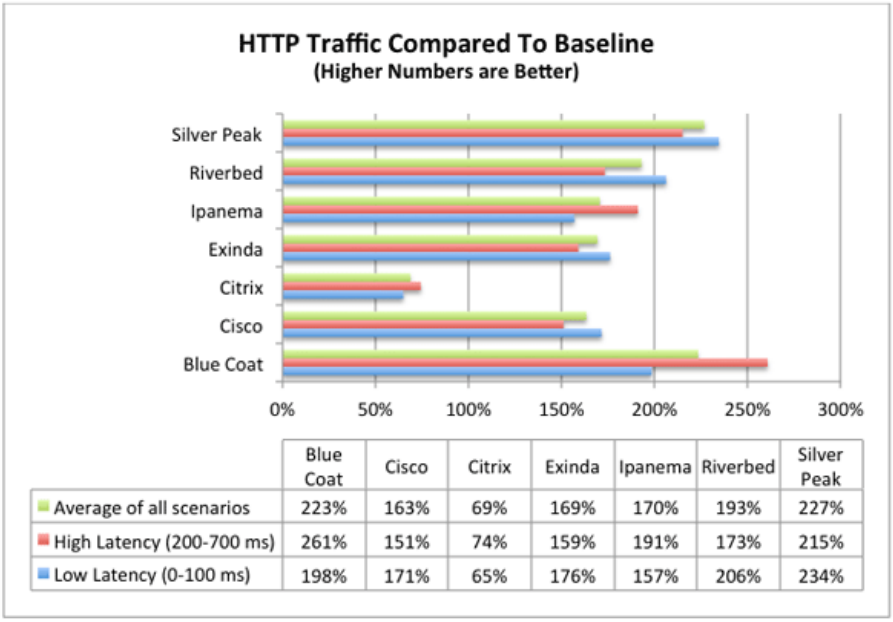
Un aspecto a considerar es que la compresión WAN pura y la deduplicación rara vez se “pagan” en áreas donde el ancho de banda es barato y abundante, es decir, gran parte de las Américas, Europa, Asia y Australia. Pero cuando las WANs son intercontinentales, los precios se disparan y el hardware de compresión a menudo puede justificarse sobre una base de costo pura, ignorando otros factores. En el estudio se obvio el intercambio de archivos CIFS, debido a la gran variación en la optimización de CIFS, y los resultados podrían resultar muy sesgados basados en pequeñas variaciones en cómo se prueba cada escenario.

Para la prueba se empezó con el tráfico HTTP y HTTPS: sitios web internos, SharePoint, aplicaciones POS y ERP basadas en web. Para ello, se comparó las transacciones que se podían completar en un circuito de 45Mbps con y sin optimización. Estas son las principales conclusiones del estudio realizado y que se analizan en el estudio de la revista NetworkWorld.

8.3. TRÁFICO HTTP

Para el tráfico HTTP puro, Blue Coat Mach5 sobresalió en las redes de alta latencia con una sorprendente mejora del 260% en las transacciones completadas en comparación con la línea de base. Silver Peak VX-series ayudó más en redes de baja latencia, mejorando el conteo de transacciones en un 234%. En general, todos los productos excepto Citrix CloudBridge lo hicieron bien, aumentando el rendimiento en al menos el 170%.

Para el tipo de tráfico que se usó en el estudio (una mezcla de objetos HTTP), se estima que la mayoría de los administradores de redes encontrarán que los usuarios finales reportan el mejor rendimiento con Silver Peak VX-series y Blue Coat Mach5 y un rendimiento ligeramente inferior, pero no significativamente diferente, entre Riverbed Steelhead, Ipanema ip | engine, Exinda x800-series, y Cisco WAAS.



	Blue Coat	Cisco	Citrix	Exinda	Ipanema	Riverbed	Silver Peak
Average of all scenarios	223%	163%	69%	169%	170%	193%	227%
High Latency (200-700 ms)	261%	151%	74%	159%	191%	173%	215%
Low Latency (0-100 ms)	198%	171%	65%	176%	157%	206%	234%

Figura 16: Comparación de tráfico HTTP *[8]*
Fuente: http://www.networkworld.com/article/2363942/lan-wan/riverbed-wins-7-vendor-wan-optimization-test.html

8.4. TRÁFICO HTTPS

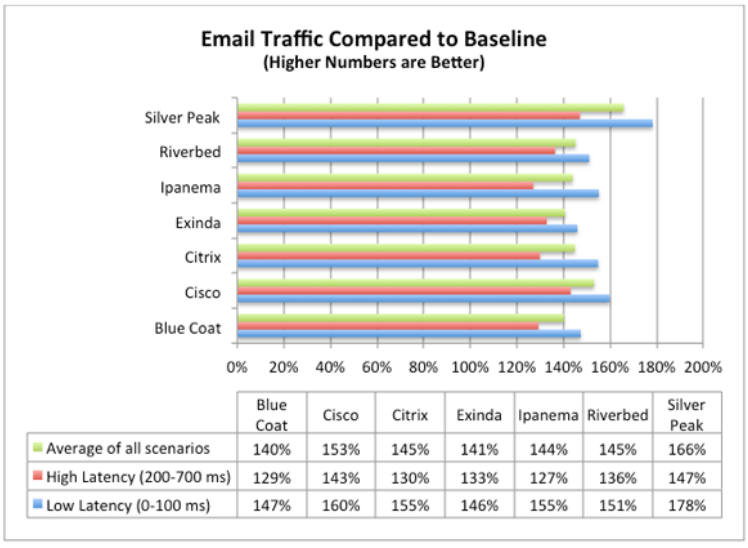
La gran disparidad entre el tráfico HTTP y el HTTPS es un factor importante para que los administradores de red consideren una solución específica. Si el tráfico es todo HTTP y HTTPS y el

rendimiento es el criterio número uno, Riverbed Steelhead ciertamente es la mejor solución para ambos protocolos, con Silver Peak VX-series como serios contendientes en este aspecto de la optimización.

- Diferentes vendedores tienen diferentes enfoques de HTTPS, sin embargo, es bastante claro que el enfoque de Blue Coat utilizado no era el adecuado, ya que cayó de ser el mejor rendimiento en las redes de baja latencia HTTP al peor rendimiento en las redes HTTPS. Se concluyó que Blue Coat Mach5 en una sesión típica de HTTPS, reducirá su rendimiento en más de la mitad.
- Riverbed fue el que mejor trabajó en las pruebas de aceleración HTTPS: un promedio de 185% de mejora en la tasa de transacción en comparación con el tráfico no optimizado.
- Cisco WAAS y Silver Peak VX-series fueron los siguientes en lista.

8.5. TRÁFICO DE CORREO ELECTRÓNICO

Cuando se probó el tráfico de correo electrónico, se encontró significativamente menos variabilidad entre productos y protocolos. Los aumentos de rendimiento oscilaron entre el 140% de la línea de base y el 166%. La serie Silver Peak VX supera la competencia en la compresión de correo electrónico, tanto en entornos de baja latencia como de alta latencia. Blue Coat, Citrix, Exinda, Ipanema y Riverbed se encuentran juntos en un rango de cerca del 5%.



	Blue Coat	Cisco	Citrix	Exinda	Ipanema	Riverbed	Silver Peak
Average of all scenarios	140%	153%	145%	141%	144%	145%	166%
High Latency (200-700 ms)	129%	143%	130%	133%	127%	136%	147%
Low Latency (0-100 ms)	147%	160%	155%	146%	155%	151%	178%

Figura 17: *Comparación de tráfico de Correo Electrónico [8]*
Fuente: *http://www.networkworld.com/article/2363942/lan-wan/riverbed-wins-7-vendor-wan-optimization-test.html*

8.6. TRÁFICO DE VOZ SOBRE IP VOIP

En el análisis del tráfico de voz (SIP-establecido mediante Voice over IP) para ver qué pasa con los dispositivos de optimización de red. En general, la respuesta fue "nada" por dos razones: el tráfico VoIP no se puede comprimir, porque el CODEC en el teléfono VoIP ya ha hecho la compresión. Además, el tráfico UDP sin conexión no cae en el modelo "bump in the wire[1]" de los productos que se probaron. Para comprimir correctamente el tráfico UDP, los dos dispositivos tendrían que tener un túnel explícito para que conocieran el tráfico, en lugar de detectar automáticamente la compresión a través de las opciones TCP.

- El modelo Silver Peak VX-serie no funciona exactamente en un modelo de *bump-in-the-wire*, por lo que fue capaz de obtener algunas ganancias de rendimiento de tráfico UDP, un promedio de 9% de mejora respecto a la línea de base.
- Riverbed Steelhead, incluso sin la configuración específica del túnel, también fue capaz de ofrecer un aumento del 7% en el rendimiento. El resto del productos tuvo un comportamiento esperado, que van desde 99% a 103%.

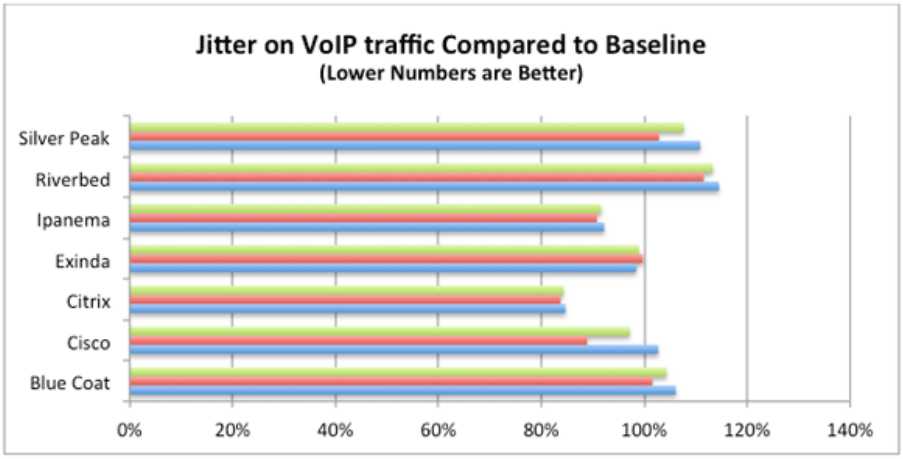


Figura 18: *Comparación de tráfico de Voz sobre IP [8]*
Fuente: *http://www.networkworld.com/article/2363942/lan-wan/riverbed-wins-7-vendor-wan-optimization-test.html*

[1] Bump in the wire: Propuesta de implementación que situa un mecanismo de seguridad de la red fuera del sistema que ha de ser protegido. Fuente: http://www.tugurium.com/gti/termino.php?Tr=bump-in-the-wire

8.7. GESTIÓN DE TRÁFICO

La ejecución de aplicaciones de una manera más rápida es el objetivo de la optimización de la red, seleccionar qué aplicaciones tienen prioridad, reservar el ancho de banda y priorizar las aplicaciones no críticas son objetivos igualmente válidos. Se agruparon esas características bajo la bandera de "gestión del tráfico".

- Ipanema ip | engine salió a la cabeza, con un sofisticado sistema de gestión de aplicaciones globales.
- Por otro lado, los motores WAVE de Cisco no tienen características de administración de tráfico incorporadas. Sin embargo, la optimización integrada de ISR de Cisco hereda todas las capacidades de gestión de tráfico de IOS.
- Los administradores de redes que piensan que la administración del tráfico es importante, pueden encontrar en la Serie x800 de Exinda, Ipanema y Riverbed Steelhead los productos con el conjunto de características más sofisticados.

Ipanema ip | engine ofrece la cartera de gestión de tráfico más innovadora según lo analizado por el estudio. La tecnología ip | motor de Ipanema ofrece una gestión global del tráfico en su herramienta de gestión Salsa. En una red tipo estrella, como sucursales agrupadas alrededor de un solo centro de datos, la gestión del tráfico es fácil porque la red es realmente una serie de líneas punto a punto, todas las cuales pueden ser controladas en ambos extremos. Sin embargo en una moderna red con múltiples centros de datos, centros de soporte, comunicaciones basadas en la nube y flujos de sucursal a rama, los circuitos virtuales punto a punto desaparecen. En su lugar, cada sitio puede estar comunicándose con varios otros sitios, con potencial para problemas de congestión

Los administradores de red también deben ser conscientes de que Ipanema ip | engine sólo soporta una arquitectura de red muy rígida y un solo modelo de gestión de tráfico. Al definir flujos de tráfico globales, Salsa gestiona el ancho de banda en función de la aplicación, no en base a cada sitio. Por lo tanto, se puede decir que "la videoconferencia obtiene una garantía de 512K por llamada" o "Citrix XenDesktop obtiene una garantía de 64K por sesión", pero no hay manera fácil de decir cuántas videollamadas o sesiones de Citrix XenDesktop pueden entrar o salir de un sitio en particular o qué porcentaje del ancho de banda de un sitio puede ser ocupado por una aplicación en particular.

Las soluciones de Riverbed Steelhead y los aparatos x800 de Exinda mostraron sofisticadas capacidades de gestión del tráfico que se clasificaron por debajo de Ipanema, pero por encima del resto. Una característica clave que destacó en ambos productos fue la identificación de la aplicación de capa 7, es decir la capacidad de aplicar reglas de gestión de tráfico basadas en la aplicación que se está utilizando y no sólo una simple configuración basada en la dirección IP y el número de puerto.

Algunas otras características se explican en el artículo, sin embargo a manera general se toma como referencia que las soluciones de optimización deben trabajar de la mano de la Gestión del tráfico para proveer mejores soluciones a sus clientes finales que en muchos casos deben afrontar problemas específicos que pueden ser atacados mediante una Gestión automatizada y altamente configurable.

8.8. VISIBILIDAD

Al contar con cientos o miles de sucursales remotas, la capacidad de analizar rápidamente y diagnosticar problemas de red y aplicaciones es una ventaja competitiva importante. La visibilidad ayuda a la resolución de problemas mediante la entrega de información que se usa para la planificación de la capacidad, análisis de tendencias y seguimiento de las incidencias. Dado que un dispositivo de optimización de red puede ver el tráfico WAN antes de obtener pasar por dispositivos que realicen tareas de VPN, NAT, etc. son un gran lugar ideal para obtener visibilidad sobre el tráfico que sale de la sucursal y es parte de la red corporativa.

Las pruebas del estudio se centraron en la evaluación de las capacidades de los productos para el análisis de tráfico a corto y largo plazo, encontrándose que Riverbed Steelhead y Exinda x800 ofrecían el conjunto de características de visibilidad más fuerte y más amplio, aunque obtener las funcionalidades completas de Riverbed Steelhead también significa incluir en el presupuesto de la solución herramientas de análisis Riverbed Cascade.

8.9. FACILIDAD DE USO

Dada la experiencia y al ser uno de los competidores con más años en el mercado, la solución de optimización WAN de Riverbed Steelhead establece el escenario para el mercado con productos que son flexibles y fáciles de usar. Los dispositivos Riverbed Steelhead no requieren configuración complicada para definir topologías de red, ejecutándose casi sin ninguna programación y gestión continua o actualización. Con Steelhead de Riverbed definiendo los estándares del mercado, casi

todos los competidores han seguido sus pasos y hacen un buen trabajo con respeto a mejorar las características de facilidad de uso. Los rezagados aquí son Silver Peak y Cisco.

Con Silver Peak VX-serie, la conexión entre dos dispositivos se gestiona de forma más directa debido a que un túnel se establece siempre entre los dispositivos. Esto significa que se tiene que poner la serie VX-fuera de cualquier Firewall, debido a que un servidor de seguridad en el lado WAN de la serie VX no se puede aplicar controles normales. Silver Peak VX-series creará automáticamente túneles si lo desea, pero entonces los túneles tienen que ser gestionados si se desea utilizar las funciones de gestión del tráfico, creando una carga adicional e innecesaria, sobre todo en una WAN de grandes proporciones.

Con respecto a Cisco el estudio muestra que tienen dos problemas importantes con Cisco WAAS y la transparencia de red, dependiendo de si se utiliza Dispositivos tipo WAVE o la funcionalidad de WASS integrada con IOS. Con los dispositivos independientes tipo WAVE, el software WAAS de Cisco es incompatible con cualquier Firewall que verifican los números de secuencia TCP. Si se requiere trabajar con el firewall, es necesario desactivar la comprobación de secuencia de número TCP (que es una función de seguridad que se puso en este tipo de equipos por una buena razón). Por otro lado, si se elige utilizar WAAS integrado en IOS, entonces se tiene un problema de transparencia diferente porque la versión integrada en IOS no puede operar de forma transparente en las redes existentes.

Otro producto que se encuentra en la parte baja de la lista es Citrix CloudBridge 2000 en el área de la facilidad de uso ya que las comunicaciones de dispositivo a dispositivo sólo se gestionan de forma automática cuando no se está usando conexiones SSL. Una vez que use SSL se requiere del administrador de red para hacer una gestión mucho más explícita y realizar configuraciones y cambios en la topología, lo que complica innecesariamente las implementaciones.

9. ELECCIÓN DE LA HERRAMIENTA DE OPTIMIZACIÓN PARA EL AMBIENTE DE LABORATORIO.

Dentro del mercado de soluciones de Optmización WAN se pueden encontrar algunas soluciones propietarias y de código abierto que prometen optimizar los enlaces WAN de los clientes. Dentro de las herramientas de código abierto se pueden encontrar muchas herramientas que se pueden implementar haciendo uso de imágenes pre-configuradas como Wanos[2] que según su página web implementa herramientas de Calidad de Servicio, Optmización Bi-direccional y retención de información duplicada y funcionalidades de Router; otras herramientas como OpenNOP[3] realizan las mismas tareas y guardan bastante similitud con respecto a soluciones propietarias como Riverbed y NetScaler SD-WAN de la empresa Citrix. Bajo este escenario se ha optado por buscar la opción de un tercero para determinar la mejor opción en soluciones de optmización, decidiendo ya de antemano que no se escogerán soluciones de código abierto debido a la falta de información estructurada sobre la implementación y configuración de las herramientas, adicionalmente inicialmente se está contemplando presentar la solución con un soporte de fábrica o del proveedor que permita tener un apoyo adicional en caso de mal funcionamiento de la aplicación, adicionalmente se ha visto que las soluciones propietarias liberan periódicamente actualizaciones de software sobre las aplicaciones que corrigen errores o vulnerabilidades de seguridad.

Para esto se ha buscado la opinión de Gartner[4], en dónde se analiza las soluciones del mercado y define el mercado de optimización WAN como soluciones que mejoran el desempeño de aplicaciones que cruzan a través de la WAN y también reduce los gastos por servicios de enlaces WAN. Según el gráfico de la Figura 9.1 se pueden identificar 3 soluciones que pertenecen al cuadrante de líderes que sobresalen y son Cisco, Riverbed y Silver Peak, adicionalmente podemos ver a Citrix en el cuadrante de Retadores. Con esta información se decidió analizar estas opciones y escoger una de estas herramientas para levantar el laboratorio de pruebas y analizar las propiedades de los optimizadores WAN.

[2] Página web: http://wanos.co/wan-optimization/
[3] Página web: http://www.opennop.org/features.php
[4] Gartner Inc. es una empresa consultora y de investigación de las tecnologías de la información con sede en Stamford, Connecticut, Estados Unidos. [Wikipedia - https://es.wikipedia.org/wiki/Gartner_(empresa)]

Figura 19: *Cuadrante de Gartner para Optimización WAN*
Fuente: *https://www.gartner.com/doc/reprints?id=1-37LNFTS&ct=160520&st=sb*

El primer ejercicio que se realizó fue el de buscar la opción de obtener licencias de prueba para la evaluación de los productos de manera gratuita. Se escogió este parámetro como primera opción ya que al tratarse de un ambiente de laboratorio es necesario contar con el producto para realizar las pruebas respectivas y si bien otras soluciones pueden proveer mejores resultados o incluso demos con equipos físicos la idea es realizar una batería de pruebas en las que se pueda observar los mecanismos de optimización. Según la Tabla 3 se puede observar que existen dos productos que pueden entregar licencias de Demostración con un tiempo mayor a 30 días.

Producto	Licencia de Demo	Tiempo
Cisco	No	N/A
Riverbed	Si	90 días
Silver Peak	Si	30 días
Citrix	Si	No especifica

Tabla 3: *Soluciones de Optimización WAN – Disponibilidad de Licencia de Demostración*

Como segundo parámetro se realizó una búsqueda de canales de distribución o representantes de las marcas en Ecuador, la búsqueda para el caso de Silver Peak no arrojó canales en el Ecuador lo

que dentro del esquema de tener un contacto cercano con la marca la elimina. El resto de marcas tienen representantes locales en varias empresas integradoras de tecnología, motivo por el cual se ha decidido centrar el análisis entre las marcas Citrix y Riverbed.

Según el análisis de Gartner la solución de Citrix se vende como parte de otras soluciones de este proveedor y no posee de un amplio canal de integradores que puedan implementar las soluciones, en el otro lado de la moneda Riverbed se señala como una solución costosa y con una política de descuentos bastante limitada. Finalmente se ha decidido que la herramienta para la optimización de canales WAN sea la del fabricante Riverbed.

10. ARQUITECTURA DE LA SOLUCIÓN DE OPTIMIZACIÓN DE ANCHO DE BANDA WAN PARA EL AMBIENTE DE LABORATORIO

El ambiente usado para el laboratorio en el que se evaluará la solución de optimización WAN es un conjunto de equipos y herramientas en las que se usa herramientas de virtualización que permiten optimizar los recursos y no necesita el despliegue de una solución extensa y complicada. Como se observa en la Figura 20 se necesitó de 3 equipos físicos en los que se ha desplegado los componentes necesarios para realizar la simulación del ambiente. En el Servidor Virtual – Vmware se instaló las máquinas virtuales que permiten la simulación del enlace WAN y de las máquinas virtuales con los equipos de la solución de Riverbed, como ese analiza en el sección 11 la herramientas necesitan de pasos de configuración y personalización en las que se puede recrear las condiciones de la Unidad Educativa "Eloy Alfaro". Es importante aclarar que al tratarse de ambiente controlado las condiciones siempre serán las ideales y no se puede comparar con un ambiente productivo real.

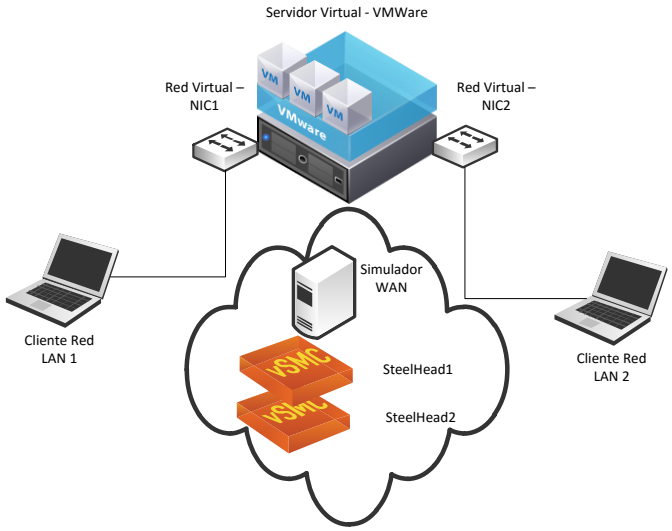


Figura 20: *Ambiente físico para Simulación*

Para la creación de las máquinas virtuales que la herramienta Riverbed necesita es importante entender los diferentes modelos que el fabricante puede entregar y los parámetros que determinan el modelo que se debe escoger, por ese motivo se ha consultado en el Documento de Especificaciones técnicas de Riverbed y se ubicó que el modelo VCX255 cumple con las necesidades de la Unidad Educativa "Eloy Alfaro" ya que puede manejar anchos de banda WAN de hasta 6 Mbps, adicionalmente los requerimientos de procesador y memoria son mínimos y pueden cumplirse perfectamente en el ambiente que se ha montado para el efecto.

Riverbed® SteelHead™

Model Specifications

SteelHead CX for Virtual

Model	VCX255				VCX555			VCX755			VCX1555		
Configurations	U	L	M	H	L	M	H	L	M	H	L	M	H
Optimized WAN Capacity	2 Mbps	6 Mbps			6 Mbps	10 Mbps		10 Mbps		20 Mbps	50 Mbps		100 Mbps[3]
Optimized TCP and UDP flows	50	75	150	230	250	400	650	900	1,500	2,300	3,000	4,500	6,000
QoS Bandwidth[1]	4 Mbps	12 Mbps			12 Mbps	20 Mbps		45 Mbps			100 Mbps		
QoS Rules/Classes[2]	No Published Limit (global QoS limit of 2000 rules and 2000 classes)												
Disk Capacity Advised	88 GB				118 GB			140 GB		188 GB	438 GB		
Max Data Store Capacity	50 GB				80 GB			102 GB		150 GB	400 GB		
Virtual CPUs Required (min)	1 x 1.0 Ghz				1 x 1.2 Ghz			2 x 1.2 Ghz			4 x 1.2 Ghz		
Reserved Memory Required	2 GB				2 GB			2 GB		4 GB	8 GB		

Figura 21: *Modelos de Equipamiento SteelHead [9]*
Fuente: *https://www.riverbed.com/document/fpo/Products/SteelHead/Spec%20Sheet%20-%20Steelhead%20Family.pdf*

Como siguiente punto para la creación del enlace WAN, se analizaron las opciones de creación de un ambiente físico mediante la conexión de dos routers de manera local y simular el enlace de comunicaciones, sin embargo al ser equipamiento con el que la Unidad Educativa "Eloy Alfaro" no podía proveerlo se optó por encontrar una herramienta de software que simule la comunicación WAN. Para esto se encontró la herramienta WANem[5] que permite simular diferentes escenarios de comunicación con la posibilidad de trabajar con parámetros como Retardo en la red, corrupción de paquetes, desconexiones, Jitter, etc.

[5] Página Web: http://wanem.sourceforge.net/

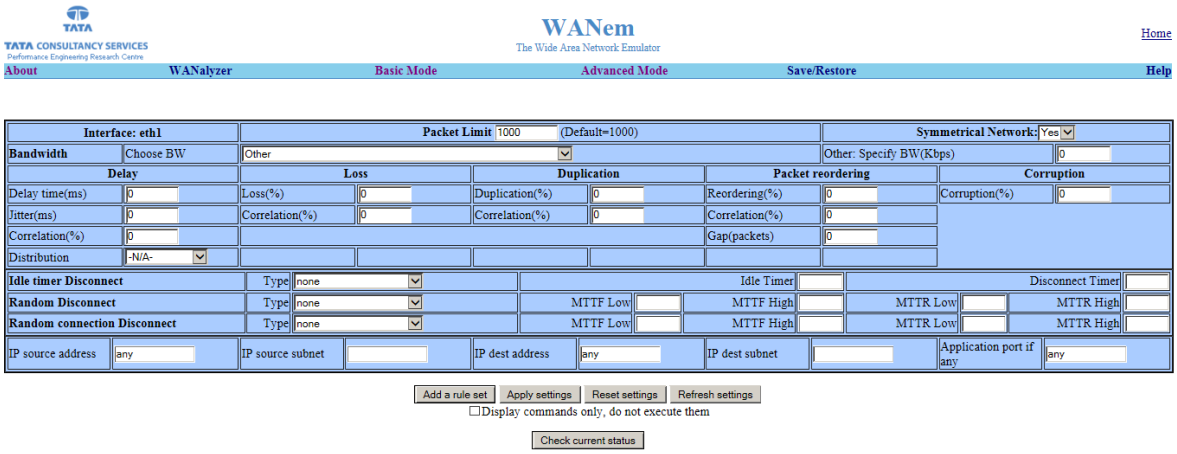


Figura 22: Opciones de Configuración de la Herramienta WANem
Fuente: Captura de Pantalla en el ambiente Virtual implementado

En la Figura 22, se puede observar varias de las opciones de la herramienta para la simulación del enlace WAN, la principal ventaja de esta herramienta es que simplifica la creación de escenarios de pruebas y limita las configuraciones de routers y switches de borde a un ambiente de virtual en el que se deben definir las direcciones IP de las interfaces LAN de las dos redes que van a simular las comunicaciones.

10.1. PARÁMETROS DE LÓGICOS DEL LABORATORIO

El ambiente de laboratorio va a simular dos redes locales que se comunican a través de una WAN, como parte de este ambiente se ha definido las siguientes redes con sus respectivas configuraciones de red que se detallan en la Tabla 4.

Característica	Red A	Red B
Red	192.168.10.0	192.168.11.0
Máscara	255.255.255.0	255.255.255.0
Puerta de Enlace	192.168.10.40	192.168.11.40
Optimizador Red A	192.168.10.10	N/A
Optimizador Red B	N/A	192.168.11.10

Tabla 4: *Soluciones de Optimización WAN – Disponibilidad de Licencia de Demostración*

11. DESARROLLO DE LA SOLUCIÓN EN LABORATORIO

Para la implementación de la solución a nivel de laboratorio se utilizaron herramientas de código libre y licencias de demo del equipo de optimización WAN. Todo esto ejecutado en un ambiente mixto de equipos virtuales y físicos en los que se realizaron diversas pruebas de transferencias de archivos, accesos web y comunicación mediante telefonía IP.

11.1. CONFIGURACIÓN DEL AMBIENTE DE LABORATORIO

Como requerimiento adicional para permitir el routeo dentro de esta herramienta se realizó una configuración a nivel del sistema operativo para habilitar la comunicación entre las dos redes implementadas.

```
exit2shell
route add -net 192.168.10.0 netmask 255.255.255.0 dev eth0
route add -net 192.168.11.0 netmask 255.255.255.0 dev eth1
wanem
```

Figura 23: *Comando para configuración ruteo*
Fuente: *https://sourceforge.net/p/wanem/discussion/711494/thread/896bf1cb/*

Para la configuración de la herramienta de optimización se usó la opción del "Virtual Appliance[6]" que corre sobre un ambiente VMware, para esto desde la página web del fabricante Riverbed es posible descargar licencias de demostración de la herramienta y realizar pruebas sobre el ambiente desplegado. No se muestra a detalle el proceso de instalación del Appliance Virtual y el proceso de licenciamiento de demo, sin embargo si se incluye el proceso de configuración inicial del equipo.

Una vez instalado el equipo es necesario realizar las configuraciones iniciales, para lo cual se ingresan los siguientes comandos desde la consola del mismo.

```
Step 1: Hostname? [SteelHead1]
Step 2: Use DHCP on primary interface? [no]
Step 3: Primary IP address? [192.168.10.20]
Step 4: Netmask? [255.255.255.0]
Step 5: Default gateway? [192.168.10.40]
Step 6: Primary DNS server?
Step 7: Domain name?
Step 8: Admin password?
Step 9: SMTP server? []
Step 10: Notification email address?
Step 11: Set the primary interface speed? [auto]
Step 12: Set the primary interface duplex? [auto]
```

[6] Virtual Appliance: equipo virtual que se instala sobre un Hypervisor para compartir los recursos físicos del mismo.

```
Step 13: Would you like to activate the in-path configuration? [yes]
Step 14: In-Path IP address? [192.168.10.10]
Step 15: In-Path Netmask? [255.255.255.0]
Step 16: In-Path Default gateway? [192.168.10.40]
Step 17: Set the in-path:LAN interface speed? [auto]
Step 18: Set the in-path:LAN interface duplex? [auto]
Step 19: Set the in-path:WAN interface speed? [auto]
Step 20: Set the in-path:WAN interface duplex? [auto]
```

Figura 24: *Configuración de Equipo SteelHead Virtual [10]*
Fuente: *Captura de pantalla de Appliance Virtual*

Posterior a esta configuración es posible ingresar a la interfaz Web de la herramienta y verificar

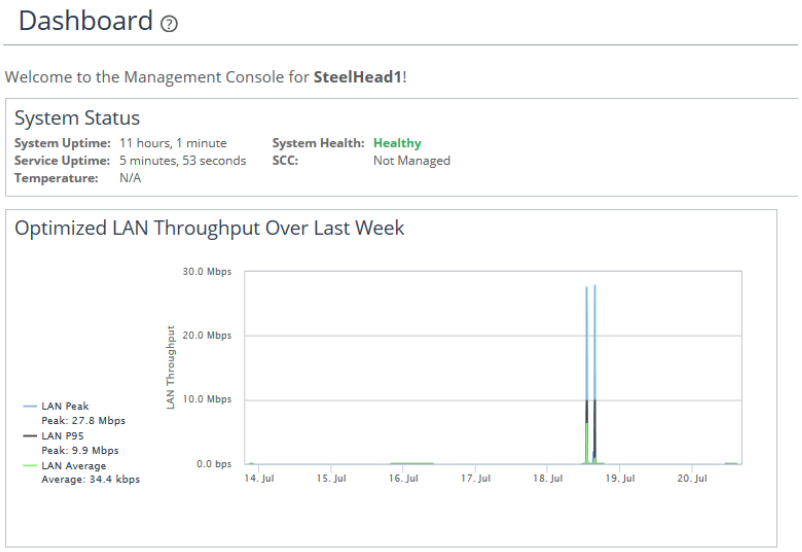


Figura 25: *Estado de Salud del Equipo Virtual*
Fuente: *Captura de pantalla de Appliance Virtual - Web*

Dentro de la herramienta se puede navegar dentro de diferentes menús en los que es posible verificar el estado de la optimización, conexiones activas y estado de las interfaces del equipo. Más adelante se presentarán los diferentes menús de la herramienta y se presentarán capturas de pantalla de los informes que presenta para cada uno de los escenarios que se han armado en el laboratorio.

11.2. ARQUITECTURA LÓGICA DEL AMBIENTE DE LABORATORIO

Para entender la problemática del cliente es necesario plantear el escenario donde se desarrollará el laboratorio, por este motivo a continuación se presenta los componentes del laboratorio,

direccionamientos IP y datos del ambiente relevantes que se han considerado durante la implementación.

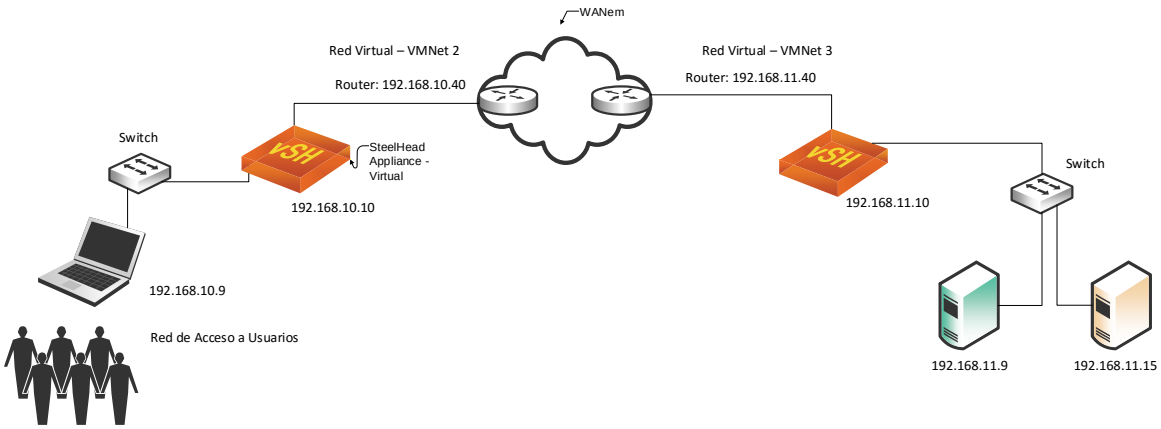


Figura 26: Diagrama de Red del Laboratorio

En este ambiente se realizarán varias pruebas de tráfico entre las redes 192.168.10.0 y la red 192.168.11.10, jugando con los parámetros de capacidad de la WAN según lo observado en la Figura 11.5. A continuación se detallan los casos de prueba que se usarán para el acceso

11.3. ESCENARIO DE LABORATORIO

11.3.1. CONSIDERACIONES DE LABORATORIO

En este escenario de laboratorio se realizará una transferencia de archivos tipo CIFS (SMBv1) entre en cliente con sistema Operativo Windows 7 y un servidor en el centro de Datos mientras se monitorea y guarda los tiempos en la transferencia. Para este escenario se usará el ambiente previamente configurado en los puntos anteriores de este documento y se buscará analizar los beneficios en la experiencia del usuario en la transferencia de archivos. Para este escenario se considerará una latencia de 50ms y un ancho de canal de 1.544 Mbps.

Como primer paso se realizará un mapeo de la unidad del servidor en la ubicación remota en nuestro cliente Windows y se realizará una prueba de conectividad entre los equipos de origen y destino, adicionalmente se verifica que el tiempo configurado de latencia corresponda al escenario propuesto.

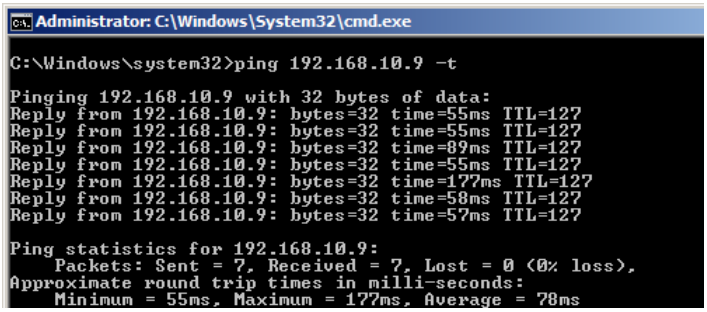


Figura 27: *Prueba de Conectividad*

Para verificar la capacidad configurada del canal es la correcta se realiza una prueba de canal mediante la herramienta Iperf[7] que nos permite medir el ancho de banda del canal, para esto se debe instalar en el origen y en el destino la herramienta y correr un set de pruebas que determinan la capacidad del canal. En la Figura 11.7 se observa el resultado de la prueba en el que se confirma que la capacidad del canal es de 1 T1 (1.544Mbps).

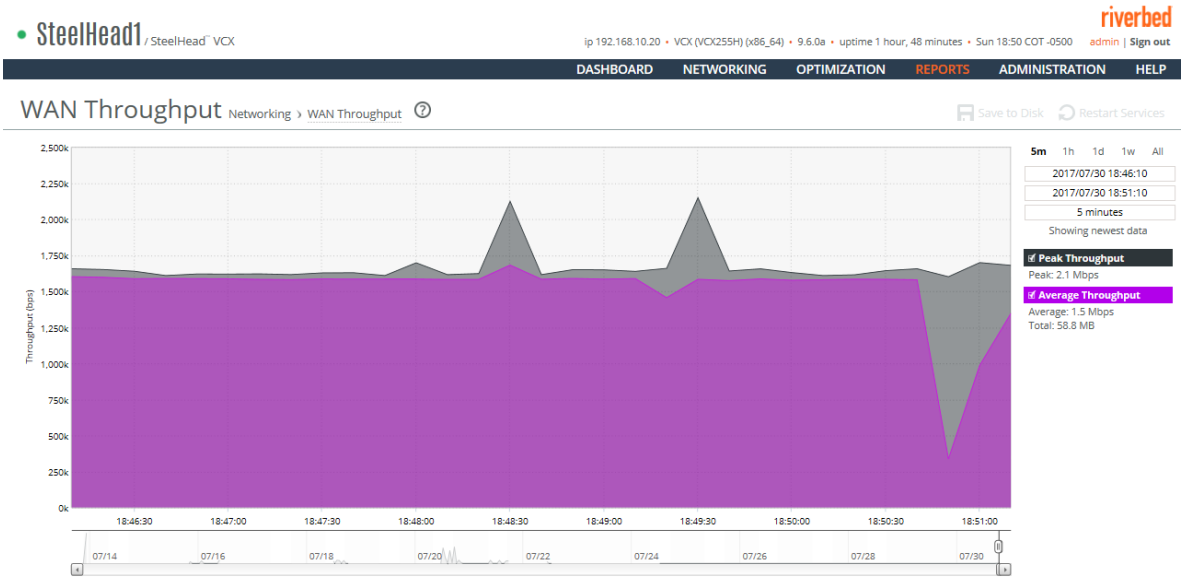


Figura 28: Verificación de la velocidad del Canal dentro de la consola SteelHead

Para la configuración del mapeo de la unidad se debe definir previamente los recursos compartidos, en este caso se ha escogido un repositorio de información varia de documentos e imágenes. Una vez configurada esta opción es posible realizar una verificación entre los equipos optimizadores en la opción web de la herramienta para esto se debe ingresar en la opción de Reports – Optmization

[7] Página web: https://iperf.fr/

– Peers; esto se lo puede observar en la Figura 29 con esto se puede identificar que los equipos se han identificado y que existe comunicación entre ambos.

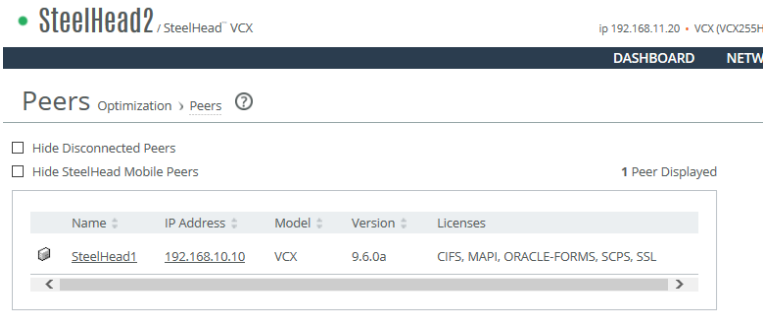


Figura 29: *Verificación de conectividad entre equipos*

Otro reporte importante que es necesario verificar es el reporte de conectividad que se encuentra en Reports – Optmization – Current Connections en donde se puede verificar las conexiones que están activas y nos enseña un valor de la reducción del envío de información a través de la WAN. Estas verificaciones permiten asegurarse que los equipos que forman parte de la solución están trabajando apropiadamente y que la solución va a trabajar adecuadamente en el ambiente de laboratorio que se ha planteado.

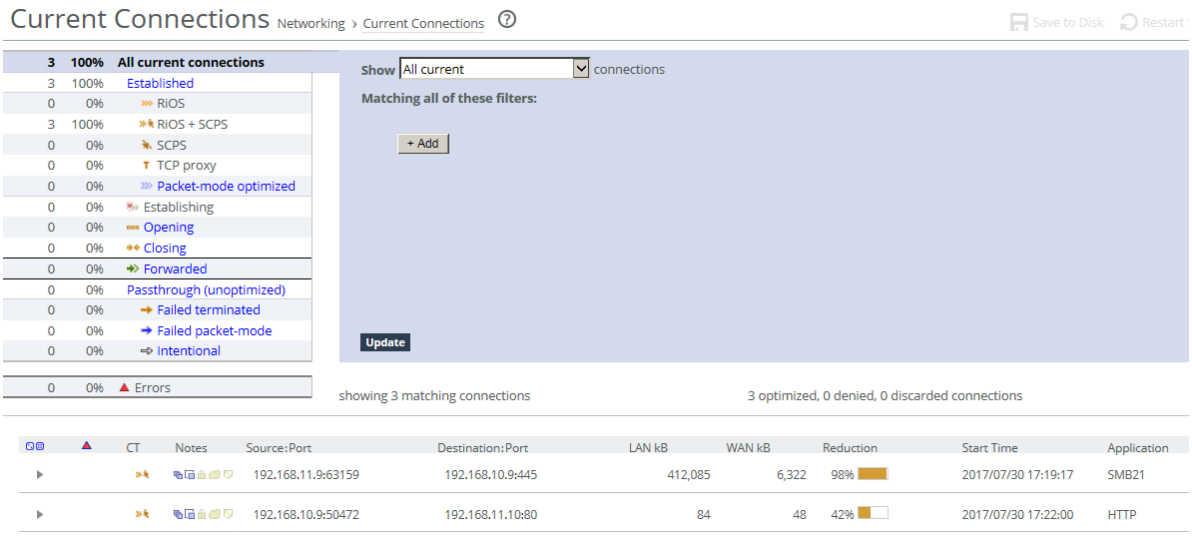


Figura 30: *Informe de "Current Connections"*

11.3.2. TRANSFERENCIA DE ARCHIVOS EN FRÍO MEDIANTE CIFS/SMBV1

Como paso inicial para realizar el proceso de medición de la optimización se realiza un reinicio de los servicios de y un borrado del caché del equipo, como se muestra en la Figura 31, este procedimiento nos garantiza que los datos y mediciones de optimización nos puedan entregar información sobre el comportamiento del equipo.

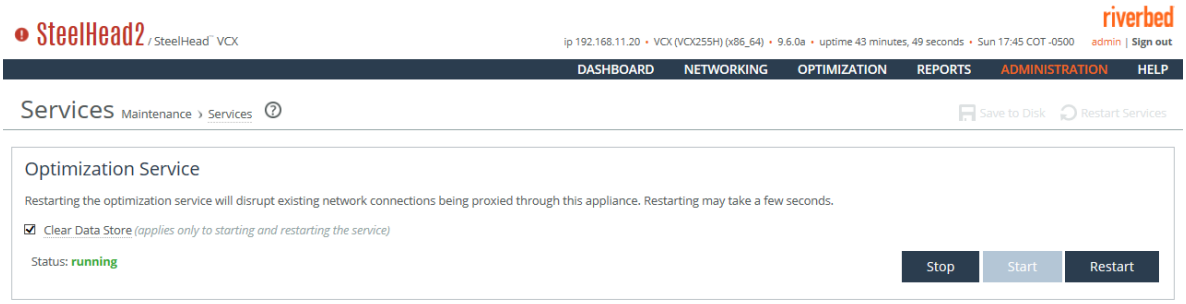


Figura 31: *Reinicio de servicios de Optimización*

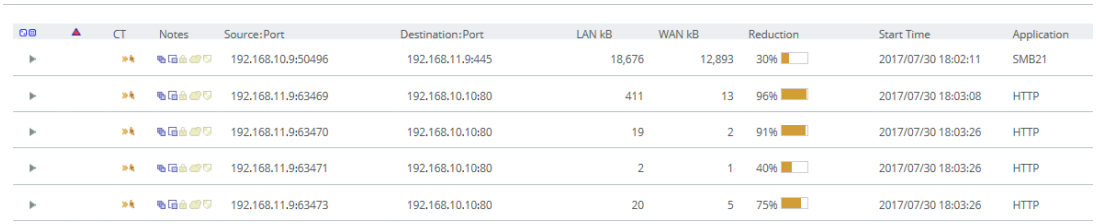
En la primera prueba se va a realizar una transferencia de un archivo de Power Point con un peso 6,798 KBytes. Como se revisó en los capítulos previos el equipo de optimización maneja el concepto de transferencia en frío y transferencia en caliente, en el cual se explica que la primera transferencia o envío inicial en frío de un mismo archivo va a durar un mayor tiempo que transferencias futuras de archivos iguales o envíos en caliente. Para medir la velocidad de la trasferencia se han realizado varias pruebas y se ha cronometrado el tiempo que toma cada una de las pruebas como se observa en la Tabla 5.

	Tamaño del Archivo	Tiempo Cronometrado [s]
Prueba 1 – Frío	6798 Bytes	**38,56**
Prueba 2 – Caliente	6798 Bytes	**36,91**
Prueba 3 – Caliente	6798 Bytes	**36,98**
Prueba 4 – Caliente	6798 Bytes	**36,70**
Prueba 5 – Caliente	6798 Bytes	**37,10**

Tabla 5: *Pruebas de envío de información*

En esta prueba de envío de la información se puede ver que la carga inicial se demoró cerca de 1,5 segundos que las descargas posteriores. Si bien el tiempo parece no ser muy significativo es importante analizar que al momento se trabaja en un ambiente en el que se analiza una única

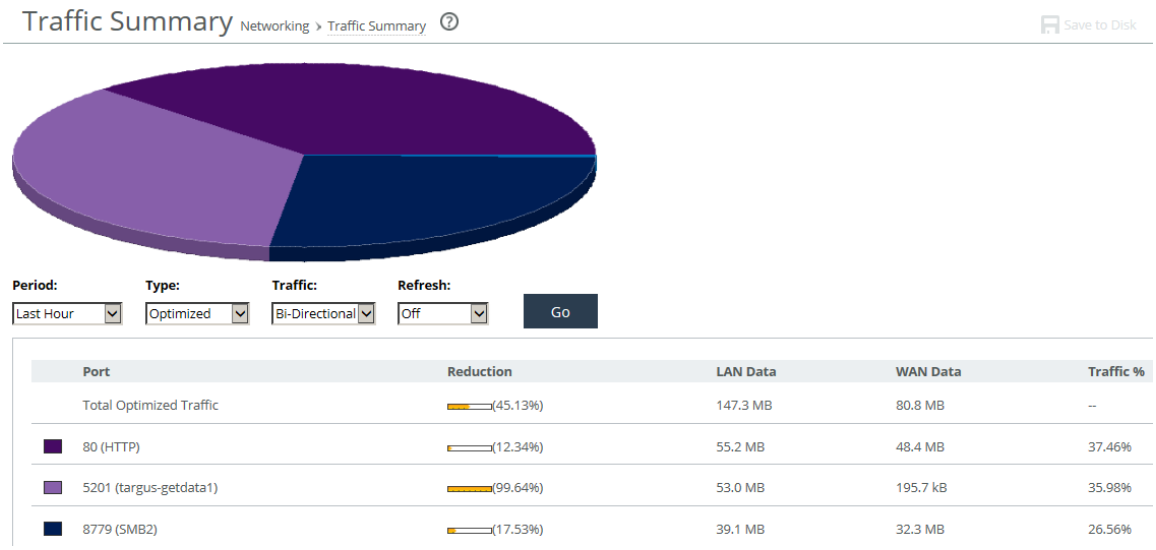
conexión de cliente hacia un servidor de archivos, este ahorro en el tiempo de descarga se verá multiplicado en el ambiente de la Unidad Educativa “Eloy Alfaro” que cuenta con varios clientes que acceden a los recursos compartidos de la institución. En la Figura 32 se puede observar una captura de la herramienta que la que se observa las conexiones existentes y el tipo de aplicación que está haciendo uso del canal WAN, específicamente se observa que el para la aplicación SMB21 (CIFS) existe una tasa de reducción aproximada del 30% en el envío de información.



CT	Notes	Source:Port	Destination:Port	LAN kB	WAN kB	Reduction	Start Time	Application
		192.168.10.9:50496	192.168.11.9:445	18,676	12,893	30%	2017/07/30 18:02:11	SMB21
		192.168.11.9:63469	192.168.10.10:80	411	13	96%	2017/07/30 18:03:08	HTTP
		192.168.11.9:63470	192.168.10.10:80	19	2	91%	2017/07/30 18:03:26	HTTP
		192.168.11.9:63471	192.168.10.10:80	2	1	40%	2017/07/30 18:03:26	HTTP
		192.168.11.9:63473	192.168.10.10:80	20	5	75%	2017/07/30 18:03:26	HTTP

Figura 32: *Información de Optimización de conexiones*

Dentro de los reportes que la herramienta nos entrega se puede analizar un reporte que presenta la información consolidada de la cantidad de tráfico que se ha optimizado. En la Figura 33 se observa que esta optimización para el tráfico SMB2 ó CIFS tiene una optimización de cerca del 27%. Es importante que la comunicación entre las dos redes que forman parte del laboratorio al momento de las pruebas generó cerca de 150 MBytes de información, sin embargo por la red WAN viajó una reducida cantidad de 80 MBytes de información



Port	Reduction	LAN Data	WAN Data	Traffic %
Total Optimized Traffic	(45.13%)	147.3 MB	80.8 MB	--
80 (HTTP)	(12.34%)	55.2 MB	48.4 MB	37.46%
5201 (targus-getdata1)	(99.64%)	53.0 MB	195.7 kB	35.98%
8779 (SMB2)	(17.53%)	39.1 MB	32.3 MB	26.56%

Figura 33: *Información de tráfico optimizado*

11.3.3. DESCARGA DE INFORMACIÓN DESDE UN SERVIDOR HTTP

Como segunda prueba sobre el ambiente de laboratorio ser realizó una descarga de información desde un servidor HTTP de archivos, esta descarga fue de aproximadamente de 50 MBytes de información y según lo que se observa en la Figura 34 tuvo una optimización de alrededor del 37%. En este experimento se cronometraron los tiempos de descarga siendo estos bastante similares a los de la Tabla 5. Bajo estos escenarios se tomó un reporte que se observa en la Figura 34 en la que se puede ir revisando en tiempo real la optimización medida en Reducción de información (*Data Reduction*), que para el experimento en conjunto tuvo un promedio de 7.6%. Al final de estos experimentos se pudo evidenciar que la reducción individual puede llegar a no ser significativa pero en conjunto entre varios usuarios puede representar un gran ahorro en el consumo total del canal de datos WAN.

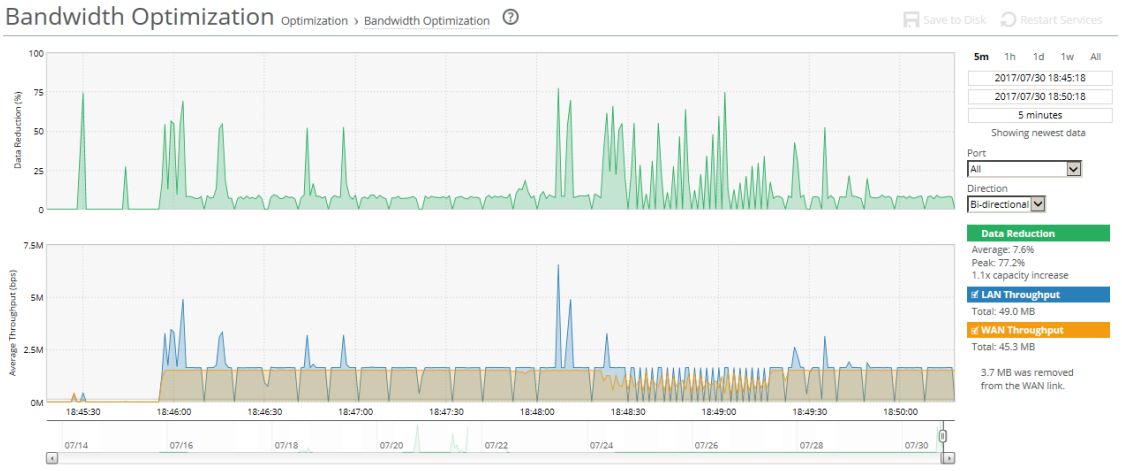


Figura 34: *Optimización de Canal*

11.3.4. ACCESO MÚLTIPLE DE USUARIOS A UNA PÁGINA WEB DE LA INTRANET

Considerando un escenario similar al del punto 11.3.1 se realizó una prueba de acceso múltiple hacia una página web montada sobre un servidor Apache Tomcat y se realizó el solicitudes simultáneas para verificar la optimización WAN que puede realizar los equipos optimizadores WAN. Para realizar esta prueba se utilizó la herramienta Apache JMeter[8] que es una herramienta Open Source que permite realizar pruebas de carga sobre un servidor web. Bajo este escenario se configuraron diez direcciones IP desde un host cliente y se realizó las configuraciones en la herramienta para realizar la solicitud sobre el puerto 80 como se muestra en la Figura 35.

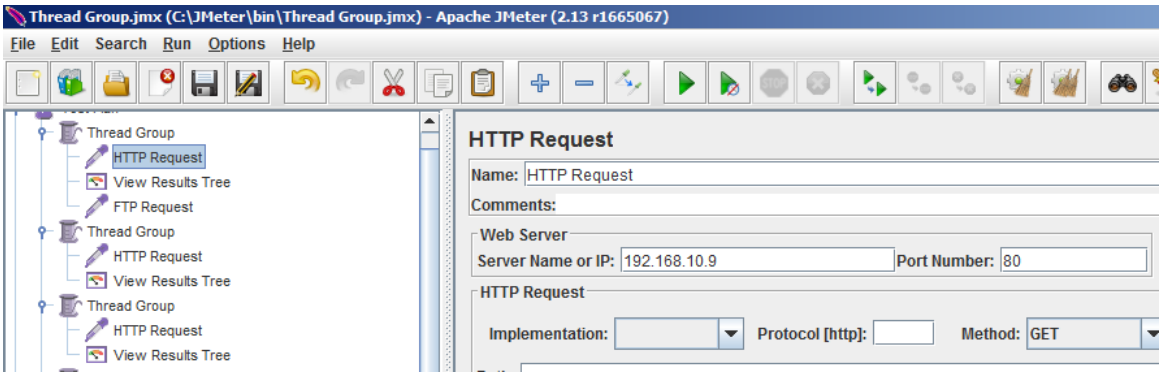


Figura 35: *Configuración de Apache JMeter*

Adicionalmente se configuró un servidor FTP sobre el mismo servidor y se realizó una prueba múltiple de acceso hacia el servidor, según lo que se observa en la Figura 36, en esta prueba no se realizó una descarga de la información debido a que se quería probar la optimización web en un ambiente con tráficos simultáneos de diferentes protocolos.

[8] Página web: http://jmeter.apache.org/

Source:Port	Destination:Port	LAN kB	WAN kB	Reduction	Start Time	Application
192.168.11.32:18474	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.32:18749	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.33:18623	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.37:18482	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.33:18629	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.37:18751	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.32:18630	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.35:18484	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.39:18753	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.37:18487	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.33:18754	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.33:18631	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.39:18501	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.32:18632	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.33:18755	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.32:18757	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.39:18633	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP
192.168.11.37:18505	192.168.10.9:80	1	0	89%	2017/08/14 17:53:51	HTTP

Figura 36: *Captura de pantalla de la optimización WAN de la herramienta Riverbed*

Según lo observado en la Figura 36, la herramienta de optimización WAN, presenta valores de reducción de información cercanos al 89% para el protocolo HTTP, eso presupone una optimización considerable para ambientes en que múltiples usuarios acceden a un recurso alojado en un servidor web.

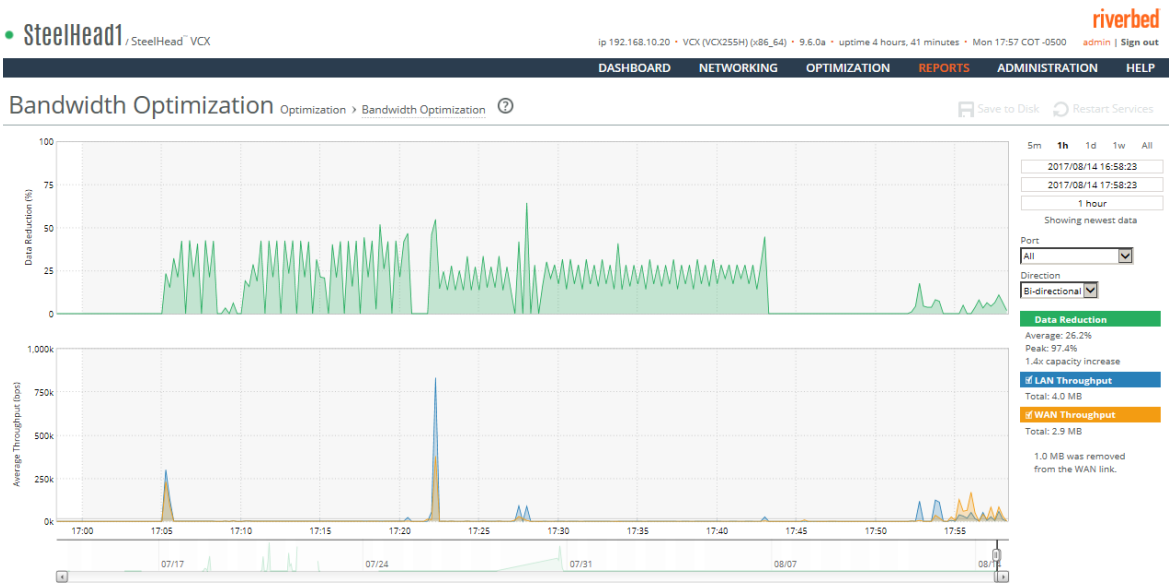


Figura 37: *Optimización de ancho de Banda en prueba de usuarios múltiples*

En la Figura 37 se puede observar que según la captura de la herramienta en el reporte de Optimización de Ancho de Banda (*Bandwidth Optimization*) durante la prueba de accesos múltiples al servidor web existió una reducción de tráfico WAN del alrededor del 26.6%. Adicionalmente el reporte entrega un aproximado del rendimiento del enlace WAN, siendo el promedio para este escenario de alrededor de 2,9 MBytes. Es importante contrastar los resultados con los de la Figura 34 en donde se obtuvo una reducción de 7,6%, en este caso se puede evidenciar como la herramienta al momento de tener mayor tráfico optimiza de mejor manera las conexiones mejora el rendimiento del enlace WAN.

11.4. ANÁLISIS DE RESULTADOS

De las pruebas realizadas se observa según la Tabla 5 que las mejoras en los tiempos de descarga son mínimas con respecto al tiempo inicial de descarga, si bien esta mejora puede parecer insignificante se debe considerar que un ambiente real se cuenta con una cantidad de usuarios del orden decenas o cientos que pueden estar trabajando simultáneamente y acceden a los recursos de la WAN de forma constante. En las Figura 32 se observa que el tráfico HTTP es altamente reducido, lo que en un escenario empresarial representa un importante ahorro en los recursos de la WAN, esto coincide con lo expuesto en los capítulos anteriores donde se explicaba que el tráfico

HTTP cuando se accede constantemente a una misma URL puede hacer uso de la memoria en caché para ahorra solicitudes que viajen a través de la WAN.

Finalmente en la Figura 32 se puede observar que el ejercicio de transferencia de archivos mediante el protocolo CIFS (SMB21), la tasa de reducción es cercana al 98%, como el caso anterior de tráfico HTTP este resultado coincide con lo expuesto en los capítulos de este trabajo que hablan de la utilización de las funcionalidad de caché para optimizar la entrega de información a clientes que deben hacer uso de la WAN para transferir archivos. Si bien la herramienta de Riverbed no permite analizar o entrega información sobre el proceso de optimización se puede deducir que las funcionalidades son de cierta manera genéricas en este tipo de soluciones y dependerá de la experiencia y desarrollo de cada fabricante para entregar una solución que provea mejores resultados.

Finalmente en el literal 11.3.4. se pudo analizar que a mayores conexiones HTTP la herramienta realiza tareas de optimización bastante altas y con valores en la prueba cercanas al 90%, esto es una gran ventaja al momento de tener un escenario real con usuarios distribuidos que acceden simultáneamente a

12. ANÁLISIS DE COSTOS DE LA SOLUCIÓN DE INFRAESTRUCTURA OPTIMIZACIÓN DE ANCHO DE BANDA WAN DE RIVERBED STEELHEAD

Como parte del proceso de análisis de la herramienta es necesario determinar en base a los costos de la solución si esta es viable para el ambiente de la Unidad Educativa "Eloy Alfaro" y los ahorros que se podrían tener al implementar la solución de Optimización WAN que se ha discutido en este trabajo. Según el acuerdo que mantiene la Unidad Educativa "Eloy Alfaro" actualmente se cuenta con un enlace de 1.544 Mbps al que se añade un valor por el arrendamiento del equipamiento de red necesario para operar el enlace, este valor actualmente asciende al valor de 380 dólares americanos al mes más el valor del IVA.

Después de realizar la consulta respectiva con el proveedor del enlace de datos se obtuvo un valor referencial de incremento del enlace WAN a 3Mbps con un costo mensual de 550 dólares americanos. En la información acerca del incremento de capacidad se detallan costos adicionales como son el cambio de equipos terminales y una nueva acometida de Fibra Óptica que suma una cantidad única de 250 dólares americanos adicionales que no se incluye en el análisis financiero de la solución pero se lo menciona como parte de los requerimientos para la ampliación de la capacidad WAN. Haciendo una estimación en base al costo fijo y proyectado a tres y cinco años se presentan los valores que se invierten en la Tabla 6 en los períodos descritos.

	Proyectado a 3 años (36 meses) – USD	Proyectado a 5 años (60 meses) – USD
Enlace WAN de 1,5Mbps	13680	22800
Enlace WAN de 3Mbps	19800	33000
Costo incremento	6120	10200

Tabla 6: *Costo del Enlace proyectado a 3 y 5 años*

De la Tabla 6 se puede observar que el incremento en capacidad representa un aumento del costo de 6120 y 10200 dólares americanos para las opciones de 3 y 5 años. Bajo este escenario es necesario comparar este valor del incremento con el costo de la implementación de una solución de Riverbed Steelhead, en la Tabla 7 y Tabla 8 se puede observar el valor de una solución de optimización que incluye los 2 equipos necesarios y los servicios de implementación y configuración incluidos como parte del servicio.

	Valor Unitario	Valor Total
SteelHead CX 255 Appliance B020 with RiOS (TAA Compliant)	1545,50	3090,99
License SteelHead CX 255 Low Appliance, 6Mbps,75 conn	898,80	1797,60
SteelHead CX Appliance 255 Gold Plus Support, 3 YEAR + 3 MESES	1419,65	2839,30
Costo de Implementación		1500,00
Total		9227,89

Tabla 7: *Costo del Enlace proyectado a 3 años*

	Valor Unitario	Valor Total
SteelHead CX 255 Appliance B020 with RiOS (TAA Compliant)	1545,50	3090,99
License SteelHead CX 255 Low Appliance, 6Mbps,75 conn	898,80	1797,60
SteelHead CX Appliance 255 Gold Plus Support, 5 YEAR + 3 MESES	2366,09	4732,18
Costo de Implementación		1500,00
Total		11120,77

Tabla 8: *Costo del Enlace proyectado a 5 años*

Comparando los valores de la solución para los escenarios de 3 y 5 años se puede observar que en el período de 5 años se estaría gastando casi el valor equivalente al de adquirir la solución. Si bien existe una diferencia en favor de la opción de incrementar los enlaces de datos y la opción de optimización WAN habría completado 5 años de utilización y contablemente se la podría dar de baja es necesario recordar que estas soluciones pueden trabajar por mucho más tiempo y pueden seguir aportando un ahorro a la institución durante algunos años adicionales.

Finalmente existen aspectos que no se ven directamente en un análisis de costos como la mejora de la experiencia del usuario al poder acceder de mejor manera a los servicios sin tiempos de espera que crean frustración en los mismos, adicionalmente la adquisición de soluciones tecnológicas involucra aspectos de capacitación e investigación en las personas encargadas de la solución que motivan al personal de TI y logran un compromiso más cerca con la institución.

13. CONCLUSIONES Y RECOMENDACIONES

- Según lo visto en el Capítulo que se describe el estado del arte de las soluciones de Optimización WAN se pudo observar que estas soluciones hoy en día están ampliamente desarrolladas por diversos fabricantes y se pueden implementar en ambientes empresariales y de medianas y pequeñas empresas que requieren optimizar sus recursos actuales debido a la constante demanda de recursos desde oficinas y/o sucursales remotas.
- Dentro de las soluciones de Optimización WAN se ha analizado que se deben superar determinados obstáculos como son la confluencia de diferentes tipos de tráficos y protocolos que compiten por los recursos de la red. Por este motivo se ha visto que la mejor estrategia previa a la implementación de este tipo de soluciones es realizar una evaluación de las necesidades actuales y verificar si en base a las capacidades de optimización de los equipos se va a tener resultados positivos al momento de la implementación.
- El principal problema que enfrentan las organizaciones al momento de analizar sus redes corporativas es la disparidad entre la capacidad de transferencia de información de la LAN y la WAN, debido al crecimiento disparejo entre ambas. En este escenario es importante analizar los requerimientos de las aplicaciones que viajan a través de la WAN y determinar la mejor solución que vayan a entregar un real incremento en el desempeño de la red WAN.
- Dentro del análisis de las soluciones existentes en el mercado, se observó que estas tienen bastantes similitudes y si bien en algunos aspectos algunas soluciones son superiores a otras la facilidad de uso juega un papel importante al momento de decidirse por alguna de ellas ya que generalmente los administradores de red buscan soluciones no disruptivas para sus ambientes, por este motivo y según lo analizado en este trabajo la opción de Riverbed Steelhead representó la mejor opción para realizar este Caso de Estudio.
- La solución de Riverbed si bien puede representar una opción cara para clientes del rango medio, es una opción que a la larga representa ahorros y mejora la experiencia de los usuarios que hace uso de la WAN. Adicionalmente al trabajar en entornos empresariales es una solución escalable que se adapta a las necesidades de los clientes.
- En la simulación de la solución en el ambiente de laboratorio se pudo observar específicamente que al ser un ambiente limitado a nivel de recursos se tenía problemas de lentitud asociado a procesos de escritura y lectura a disco. Haciendo una revisión de la literatura de los equipos optimizadores de la marca Riverbed se puede observar que las

soluciones buscan incorporar almacenamiento local del tipo SSD o Flash con el objetivo de mejorar las operaciones de escritura y lectura a disco.

- Como recomendaciones para trabajos futuros se puede analizar escenarios mixtos de Optimización WAN con una implementación de Calidad de Servicio y verificar el desempeño de las aplicaciones críticas del negocio en ese ambiente. Este escenario adicionalmente puede comparar estas dos técnicas de optimización y determinar la eficacia de cada una de ellas al momento de trabajar en un ambiente empresarial de recursos WAN limitados.
- La solución de Riverbed al ser una solución no disruptiva se podría implementar fuera de laboratorio localmente para verificar su funcionamiento y determinar la ganancia que da en el ambiente del cliente. Con este ejercicio se puede "vender" la solución al presentar en vivo sus funcionalidades y beneficios que pueden entregar.
- Si bien se analizó que las soluciones de optimización no pueden mejorar el tráfico para conexiones de Voz sobre IP, resultaría un excelente ejercicio realizar una implementación de una solución de Optimización WAN en ese ambiente y analizar como la mejora de otros tráficos diferente a VoIP pueden mejorar el desempeño del primero y resultar una opción viable para clientes con este tipo de infraestructura.

14. BIBLIOGRAFÍA

[1] Kansanen Maiju, “Wide Area Network Acceleration in Corporate Networks” – Master’s Thesis. Faculty of Technology Management, Lapperanta University of Technology.

[2] Gurpreet Singh, Lovepreet Singh, “WAN OPTIMIZATION – a review”, Journal of Advanced Research in Computer and Communication Engineering

[3] Página Web: http://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/TIC/2016/170125.Presentacion_Tics_2016.pdf

[4] Página Web: https://es.wikipedia.org/wiki/Optimizaci%C3%B3n_de_la_WAN

[5] Suarez Javier, “Estudio de las características, funcionamiento, ventajas y técnicas utilizadas en los optimizadores WAN” – Universidad Politécnica Salesiana Sede Cuenca. Cuenca 2012. Página Web: http://dspace.ups.edu.ec/bitstream/123456789/2146/16/UPS-CT002403.pdf

[6] Ted Grevers, Jr., Joel Christner. Application Acceleration and WAN Optimization Fundamentals. Cisco Systems, Inc. July 2008

[7] Parreño Pablo, “Estudio de la Optimización WAN (Wide Area Network) para la implementación de réplicas de centros de Datos” - Proyecto previo a la obtención del título de Ingeniero en Electrónica y Telecomunicaciones, Escuela Politécnica Nacional. Febrero 2017

[8] Página Web: http://www.networkworld.com/article/2363942/lan-wan/riverbed-wins-7-vendor-wan-optimization-test.html. Snyder Joel.

[9] Página Web: https://www.riverbed.com/mx/products/steelhead/steelhead-sd-wan.html

[10] “SteelHead Installation and Configuration Guide”, Version 9.2. Mayo 2016.